AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

UNITED STATES AIR FORCE RESEARCH INITIATION PROGRAM

CONDUCTED BY UNIVERSAL ENERGY SYSTEMS

U.E.S.

1987

TECHNICAL REPORT

VOLUME 2 OF 4

RODNEY C. DARRAH
PROGRAM DIRECTOR, UES

SUSAN K. ESPY
PROGRAM ADMINISTRATOR, UES

LT. COL. CLAUDE CAVENDER
PROGRAM MANAGER, AFOSR

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR-89-0830
6a. NAME OF PERFORMING ORGANIZATION UNIVERSAL ENERGY SYSTEMS INC. / 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research/XOT
6c. ADDRESS (City, State, and ZIP Code) 4401 Dayton Xenia Rd Dayton OH 45432	7b. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB DC 20332
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR / 8b. OFFICE SYMBOL (If applicable) XOT	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-85-C-0013
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F / PROJECT NO. 3396 / TASK NO. D5 / WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) USAF Research Initiation Program Volume 2 of 4	
12. PERSONAL AUTHOR(S) Program Director Rodney C. Darrah	
13a. TYPE OF REPORT Interim / 13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) April 1987 / 15. PAGE COUNT
16. SUPPLEMENTARY NOTATION	
17. COSATI CODES FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
19. ABSTRACT (Continue on reverse if necessary and identify by block number) (SEE REVERSE)	
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt Col Claude Cavender	22b. TELEPHONE (Include Area Code) (202) 767-4971 / 22c. OFFICE SYMBOL XOT

89 6 29 006

INTRODUCTION

Research Initiation Program - 1987

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1987 SFRP competed for funding under the 1987 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1987, following their participation in the 1987 SFRP.

Evaluation of these proposals was made by the contractor. Evaluation criteria consisted of:

1. Technical Excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the University

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1988.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. Rip proposals were submitted to the contractor by 1 November 1987. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1987 and December 1988.

Copies of the Final Reports are presented in Volumes I through III of the 1987 Research Initiation Program Report. There were a total of 83 RIP awards made under the 1987 program.

UNITED STATES AIR FORCE

1987 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-85-C-0013

RESEARCH REPORTS

VOLUME II OF IV

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

April 1989

TABLE OF CONTENTS

INTRODUCTION

Research Initiation Program - 1987

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1987 SFRP competed for funding under the 1987 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1987, following their participation in the 1987 SFRP.

Evaluation of these proposals was made by the contractor. Evaluation criteria consisted of:

1. Technical Excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the University

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1988.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. Rip proposals were submitted to the contractor by 1 November 1987. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1987 and December 1988.

Copies of the Final Reports are presented in Volumes I through III of the 1987 Research Initiation Program Report. There were a total of 83 RIP awards made under the 1987 program.

STATISTICS

Total SFRP Participants	159
Total RIP Proposals submitted by SFRP	117
Total RIP Proposals submitted by GSRP	7
Total RIP Proposals submitted	124

Total RIP's funded to SFRP	81
Total RIP's funded to GSRP	2
Total RIP's funded	83

Total RIP's Proposals submitted by HBCU's	11
Total RIP's Proposals funded to HBCU's	7

Laboratory	SFRP Participants	RIP's Submitted	RIP's Funded
AAMRL	13	12 (2 GSRP)	6
AFWAL/APL	8	6	4
ATL	9	8 (1 GSRP)	6
AEDC	6	4	3
AFWAL/AL	9	9 (1 GSRP)	5
LC	1	1	1
ESMC	1	0	0
ESD	1	1	0
ESC	8	8	6
AFWAL/FDL	9	8 (1 GSRP)	6 (1 GSRP)
FJSRL	9	5	5
AFGL	13	10 (1 GSRP)	7
HRL/OT	2	2	1
HRL/LR	3	3	2
HRL/MO	3	2	2
HRL/ID	0	0	0
LMC	3	1	0
AFWAL/ML	13	11 (1 GSRP)	6 (1 GSRP)
OEHL	5	3	3
AL	7	4	3
RADC	11	10	7
SAM	16	8	6
DEOMI	2	2	0
WL	7	6	4
Total	159	124	83

LIST OF UNIVERSITY THAT PARTICIPATED

Adelphi University	- 1	Meharry Medical College	- 1
Alabama, University of	- 1	Memphis State University	- 1
Alaska-Fairbanks, Univ. of	- 1	Metropolitan State College	- 1
Alfred University	- 1	Michigan State University	- 1
Arizona State University	- 1	Mississippi State University	- 4
Arkansas State University	- 1	Mississippi, University of	- 1
Arkansas, University of	- 1	Missouri-Kansas City, Univ.	- 1
Auburn University	- 1	Missouri-Rolla, Univ. of	- 3
Bishop College	- 1	Montana, University of	- 1
Capital University	- 1	Montclair State College	- 1
Catholic Univ. of America	- 1	Morehouse College	- 1
Cedarville College	- 1	Nazareth College	- 1
Central State University	- 1	Nebraska-Lincoln, Univ. of	- 2
Cincinnati, University of	- 5	New Mexico State University	- 1
Colorado, University of	- 2	New York State, Univ. of	- 3
Dayton, University of	- 7	N. Carolina A&T State Univ.	- 1
Dillard University	- 1	N. Carolina-Greensboro, Univ	- 1
Drury College	- 1	Northwestern University	- 1
Eastern Illinois University	- 1	Ohio State University	- 5
Eastern Kentucky University	- 1	Ohio University	- 2
Eastern New Mexico University	- 2	Oklahoma State University	- 1
Fairfield University	- 1	Oregon Institute of Tech.	- 1
Florida A&M University	- 1	Oregon State University	- 1
Florida, University of	- 2	Ouachita Baptist University	- 1
Fort Lewis College	- 1	Pace University	- 1
Gonzaga University	- 1	Pennsylvania State Univ.	- 1
Grambling State University	- 1	Point Loma College	- 1
Hampton University	- 1	Puerto Rico-Mayaguez, Univ.	- 1
Houston, University of	- 2	Purdue University	- 1
Howard University	- 1	Rochester Inst. of Tech.	- 1
Idaho, University of	- 1	Rose-Hulman Inst. of Tech.	- 2
Illinois-Chicago, Univ. of	- 2	Saint Paul's College	- 1
Indiana University	- 1	San Francisco State Univ.	- 1
Indiana Univ. of Pennsylvania	- 1	South Dakota State Univ.	- 1
Iowa, University of	- 1	South Florida, University of	- 2
Jackson State University	- 1	Southeastern Mass. Univ.	- 2
Jarvis Christian College	- 1	Southern Illinois University	- 2
Jesm Baromedical Res. Inst.	- 1	Southern Mississippi, Univ.	- 1
John Hopkins Evening College	- 1	Southern University	- 2
Kansas State University	- 1	St. Louis University	- 1
Kansas, University of	- 1	St. Mary's University	- 1
Kentucky, University of	- 1	Talladega College	- 1

Continued

LIST OF UNIVERSITY THAT PARTICIPATED
Continued

Lock Haven Univ. of Pennsyl.	- 1	Taylor University	- 1
Long Island University	- 1	Temple University	- 1
Louisiana State University	- 1	Tennessee Technical Univ.	- 1
Louisiana Tech. University	- 1	Tennessee, University of	- 1
Lowell, University of	- 4	Texas A&M University	- 2
Texas Southern University	- 3	Wichita State University	- 2
Texas Technical University	- 2	Wilberforce University	- 1
Texas-Austin, University of	- 1	Wisconsin-Eau Claire Univ.	- 2
Tuskegee University	- 1	Wisconsin-Madison, Univ. of	- 1
Utah State University	- 1	Wisconsin-Whitewater, Univ.	- 1
Walla Walla College	- 1	Wittenberg University	- 1
Washington State University	- 1	Worchester Polytech. Inst.	- 2
West Florida, University of	- 1	Wright State University	- 3
Western Michigan University	- 3	Xavier University	- 1

PARTICIPANTS LABORATORY ASSIGNMENT

PARTICIPANT LABORATORY ASSIGNMENT (Page 1)

AERO PROPULSION LABORATORY
(Wright-Patterson Air Force Base)

Dr. Suresh K. Aggarwal
Univ. of Illinois at Chicago
Specialty: Aerospace Engineering

Dr. Bryan R. Becker
Rose-Hulman Institute
Specialty: Engineering Science

Dr. Richard Tankin
Northwestern University
Specialty: Mechanical Engineering

Dr. Cheng-Hsiao Wu
Univ. of Missouri
Specialty: Solid State Physics

ARMAMENT LABORATORY
(Eglin Air Force Base)

Dr. Charles Bell
Arkansas State University
Specialty: Mechanical Engineering

Dr. Robert W. Courter
Louisiana State University
Specialty:Aerospace Engineering

Dr. Joseph J. Feeley
University of Idaho
Specialty:Electrical Engineering

Ms. Jennifer L. Davidson (1986), (GSRP)
University of Florida
Specialty: Mathematics

Dr. Mo Samimy (1986)
Ohio State University
Specialty: Mechanical Engineering

Dr. Elmer C. Hansen
University of Florida
Specialty:Mechanical Engineering

Dr. James Hoffmaster
Gonzaga University
Specialty:Physics

Dr. James Nail
Mississippi State Univ.
Specialty:Electrical Engineering

Dr. Meckinley Scott (1986)
University of Alabama
Specialty: Statistics

Mr. Jim S. Sirkis (1986),(GSRP)
University of Florida
Specialty: Engineering Mechanics

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
(Wright-Patterson Air Force Base)

Dr. Praphulla K. Bajpai
University of Dayton
Specialty: Immunology

Dr. Gwendolyn Howze
Texas Southern University
Specialty: Physics

Dr. Thomas Nygren
Ohio State University
Specialty: Psychology

Dr. Donald Robertson
Indiana University of PA
Specialty: Psychology

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
(Wright-Patterson Air Force Base)
(continued)

Dr. Noel Nussbaum
Wright State University
Specialty: Biology

Dr. John Westerkamp
University of Dayton
Specialty: Electrical Engineering

Dr. Jacqueline Paver (1986)
Duke University
Specialty: Biomechanical Engineering

ARNOLD ENGINEERING DEVELOPMENT CENTER
(Arnold Air Force Systems)

Dr. Suhrit K. Dey
Eastern Illinois University
Specialty: Aerospace Engineering

Dr. Surgounda Patil
Tennessee Technical University
Specialty: Math Statistics

Dr. William M. Grissom
Morehouse College
Specialty: Mechanical Engineering

ASTRONAUTICS LABORATORY
(Edwards Air Force Base)

Dr. Gurbux S. Alag
Western Michigan University
Specialty: Systems Engineering

Dr. Lawrence Schovanec
Texas Tech University
Specialty: Mathematics

Dr. John Kenney
Eastern New Mexico University
Specialty: Physical Chemistry

AVIONICS LABORATORY
(Wright-Patterson Air Force Base)

Dr. Vernon L. Bakke
University of Arkansas
Specialty: Mathematics

Dr. Narayan C. Halder
University of South Florida
Specialty: Physics

PARTICIPANT LABORATORY ASSIGNMENT (Page 3)

AVIONICS LABORATORY
(Wright-Patterson Air Force Base)
(continued)

Prof. William K. Curry
Rose-Hulman Inst. of Technology
Specialty: Computer Science

Dr. Verlynda S. Dobbs
Wright State University
Specialty: Computer Science

Dr. George W. Zobrist (1986)
University of Missouri-Rolla
Specialty: Electrical Engineering

Dr. Alastair McAulay
Wright State University
Specialty: Electrical Engineering

Dr. John Y. Cheung (1986)
University of Oklahoma
Specialty: Electrical Engineering

ENGINEERING AND SERVICES CENTER
(Tyndall Air Force Base)

Dr. William W. Bannister
University of Lowell
Specialty: Organic Chemistry

Dr. William M. Bass
The University of Tennessee
Specialty: Physical Anthropology

Dr. Peter Jeffers
S.U.N.Y.
Specialty: Chemistry

Dr. William T. Cooper (1986)
Florida State University
Specialty: Chemistry

Dr. William Schulz
Eastern Kentucky University
Specialty: Chemistry

Dr. Joseph Tedesco
Auburn University
Specialty: Civil Engineering

Dr. Dennis Truax
Mississippi State Univeristy
Specialty: Civil Engineering

Dr. Yong S. Kim (1986)
The Catholic Univ. of America
Specialty: Civil Engineering

FLIGHT DYNAMICS LABORATORY
(Wright-Patterson Air Force Base)

Mr. Thomas Enneking (GSRP)
University of Notre Dame
Specialty: Civil Engineering

Dr. Gary Slater
University of Cincinnati
Specialty: Aerospace Engineering

FLIGHT DYNAMICS LABORATORY
(Wright-Patterson Air Force Base)
(continued)

Dr. Oliver McGee
Ohio State University
Specialty: Engineering Mechanics

Dr. Shiva Singh
Univ. of Kentucky
Specialty: Mathematics

Dr. George R. Doyle (1986)
University of Dayton
Specialty: Mechanical Engineering

Dr. Forrest Thomas
University of Montana
Specialty: Chemistry

Dr. William Wolfe
Ohio State University
Specialty: Engineering

Dr. V. Dakshina Murty (1986)
University of Portland
Specialty: Engineering Mechanics

Dr. Tsun-wai G. Yip (1986)
Ohio State University
Specialty: Aeronautics-Astronautics Engineering

FRANK J. SEILER RESEARCH RESEARCH LABORATORY
(United State Air Force Academy)

Dr. Charles M. Bump
Hampton University
Specialty: Organic Chemistry

Dr. Stephen J. Gold
South Dakota State University
Specialty: Electrical Engineering

Dr. Henry Kurtz
Memphis State Univ.
Specialty: Chemistry

Dr. Howard Thompson
Purdue University
Specialty: Mechanical Engineering

Dr. Melvin Zandler
Wichita State Univ.
Specialty: Physical Chemistry

GEOPHYSICS LABORATORY
(Hanscom Air Force Base)

Dr. Lee A. Flippin
San Francisco State Univ.
Specialty: Organic Chemistry

Dr. Gandikota Rao
St. Louis University
Specialty: Meteorology

GEOPHYSICS LABORATORY
(Hanscom Air Force Base)
(continued)

Dr. Mayer Humi
WPI
Specialty: Applied Mathematics

Dr. Steven Leon
Southeastern Massachusettes
Specialty: Mathematics

Dr. Henry Nebel
Alfred University
Specialty: Physics

Dr. Timothy Su
Southeastern Massachusetts Univ.
Specialty: Physical Chemistry

Dr. Keith Walker
Point Loma College
Specialty: Physics

HUMAN RESOURCES LABORATORY
(Brooks, Williams and Wright-Patterson Air Force Base)

Dr. Patricia A. Carlson
Rose-Hulman Inst. of Technology
Specialty: Literature/Language

Dr. Ronna E. Dillon
Southern Illinois University
Specialty: Educational Psychology

Dr. Michael Matthews
Drury College
Specialty: Psychology

Dr. Stephen Loy (1986)
Iowa State University
Specialty: Management Information Sys.

Dr. Doris Walker-Dalhouse (1986)
Jackson State University
Specialty: Reading Education

Dr. John Uhlarik
Kansas State University
Specialty: Psychology

Dr. Charles Wells
University of Dayton
Specialty: Management Science

Dr. Charles Lance (1986)
University of Georgia
Specialty: Psychology

Dr. Jorge Mendoza
Texas A&M University
Specialty: Psychology

Dr. Billy Wooten (1986)
Brown University
Specialty: Philosophy, Psychology

LOGISTICS COMMAND
(Wright-Patterson Air Force Base)

Dr. Howard Weiss
Specialty: Industrial Engineering
Temple University

MATERIALS LABORATORY
(Wright-Patterson Air Force Base)

Dr. Bruce A. DeVantier
S. Illinois University
Specialty: Civil Engineering

Dr. Ravinder Diwan
Southern University
Specialty: Metallurgy

Dr. Bruce A. Craver
University of Dayton
Specialty: Physics

Dr. Robert Patsiga (1986)
Indiana Univ. of Pennsylvania
Specialty: Organic Polymer Chemistry

Dr. Gopal M. Mehrotra (1986)
Wright State University
Specialty: Metallurgy

Dr. John W. Gilmer
Penn State University
Specialty: Physical Chemistry

Dr. Gordon Johnson
Walla Walla College
Specialty: Electrical Engineering

Mr. John Usher (GSRP)
Louisiana State University
Specialty: Chemical Engineering

Dr. Nisar Shaikh (1986)
University of Nebraska-Lincoln
Specialty: Applied Mathematics

OCCUPATIONAL AND ENVIRONMENT HEALTH LABORATORY
(Brooks Air Force Base)

Dr. Richard H. Brown
Ouachita Baptist University
Specialty: Physiology

Dr. Elvis E. Deal
University of Houston
Specialty: Industrial Engineering

Dr. Kiah Edwards
Texas Southern University
Specialty: Molecular Biology

Dr. Ralph J. Rascati (1986)
Kennesaw College
Specialty: Biochemistry

ROME AIR DEVELOPMENT CENTER
(Griffis Air Force Base)

Prof. Beryl L. Barber
Oregon Institute of Technology
Specialty: Electrical Engineering

Dr. Kevin Bowyer
University of South Florida
Specialty: Computer Science

Dr. Ronald V. Canfield
Utah State University
Specialty: Statistics

Dr. Lionel R. Friedman
Worcester Polytechnic Inst.
Specialty: Physics

Dr. John M. Jobe (1986)
Miami University of Ohio
Specialty: Statistics

Dr. Louis Johnson
Oklahoma State Univ.
Specialty: Electrical Engineering

Dr. Panapkkam Ramamoorthy
University of Cincinnati
Specialty: Electrical Engineering

Dr. David Sumberg
Rochester Institute of Tech.
Specialty: Physics

Dr. Donald Hanson (1986)
University of Mississippi
Specialty: Electrical Engineering

Dr. Stephen T. Welstead (1986)
University of Alabama in Hunts.
Specialty: Applied Mathematics

SCHOOL OF AEROSPACE MEDICINE
(Brooks Air Force Base)

Prof. Phillip A. Bishop
University of Alabama
Specialty: Exercise Physiology

Dr. Mohammed Maleque
Meharry Medical College
Specialty: Pharmacology

Dr. Kurt Oughstun
University of Wisconsin
Specialty: Optical Sciences

Dr. Hoffman H. Chen (1986)
Grambling State University
Specialty: Mechanical Engineering

Dr. Ralph Peters
Wichita State University
Specialty: Zoology

Dr. Stephen Pruett
Mississippi State University
Specialty: Immunology

Dr. Wesley Tanaka
University of Wisconsin
Specialty: Biochemistry

Dr. Vito DelVecchio (1986)
University of Scranton
Specialty: Biochemistry, Genetics

PARTICIPANT LABORATORY ASSIGNMENT (Page 8)

WEAPONS LABORATORY
(Kirtland Air Force Base)

Dr. Jerome Knopp
University of Missouri
Specialty: Electrical Engineering

Dr. Barry McConnell
Florida A&M University
Specialty: Computer Science

Dr. Martin A. Shadday, Jr. (1986)
University of South Carolina
Specialty: Mechanical Engineering

Dr. Randall Peters
Texas Tech University
Specialty: Physics

Dr. William Wheless
New Mexico State University
Specialty: Electrical Engineering

RESEARCH REPORTS

MINI-GRANT RESEARCH REPORTS

1987 RESEARCH INITIATION PROGRAM

Technical Report Number	Title and Mini-Grant No.	Professor
Volume I		
Armament Laboratory		
1	Report Not Available at this Time 760-7MG-025	Dr. Charles Bell
2	Effects of Bending Flexibility on the Aerodynamic Characteristics of Slender Cylinders Determined from Free-Flight Ballistic Data 760-7MG-018	Dr. Robert W. Courter
3	Image Complexity Measures and Edge Detection 760-6MG-024	Ms. Jennifer L. Davidson (1986 GSRP)
4	Report Not Available at this Time 760-7MG-070	Dr. Joesph J. Feeley
5	Advanced Gun Gas Diversion 760-7MG-012	Dr. Elmer Hansen
6	A Physical and Numerical Study of Pressure Attenuation in Solids 760-7MG-002	Dr. James Hoffmaster
7	Pyroelectric Sensing for Potential Multi-Mode Use 760-7MG-026	Dr. James Nail
8	Gaseous Fuel Injection and Mixing in a Supersonic Combustor 760-6MG-059	Dr. Mo Samimy (1986)
9	Systems Effectiveness for Targets with Repair or Replacement Facilities of Damaged Components 760-6MG-025	Dr. Meckinley Scott (1986)
10	A Pattern Recognition Application in Elastic-Plastic Boundary Element, Hybrid Stress Analysis 760-6MG-142	Mr. Jim S. Sirkis (1986 GSRP)

35	Comparison of SSM/I Rainrates and Surface Winds with the Corresponding Conventional Data in the North West Pacific Typhoons 760-7MG-072	Dr. Gandikota Rao
36	Report Not Available at this Time 760-7MG-040	Dr. Timothy Su
37	Development of a System for the Measurement of Electron Excitation Cross Sections of Atoms and Molecules in the Near Infrared 760-7MG-074	Dr. Keith Walker

Rome Air Development Center

38	Superconductor Testing 760-7MG-103	Prof. Beryl L. Barber
39	A Form and Function Knowledge Representation for Reasoning about Classes and Instances of Objects 760-7MG-003	Dr. Kevin Bowyer
40	Development and Evaluation of a Bayesian Test for System Testability 760-7MG-032	Dr. Ronald V. Canfield
41	Crystalline Silicon Electro-Optic Waveguides 760-7MG-040	Dr. Lionel R. Friedman
42	Measurements of a Slot Antenna Fed by Coplanar Waveguide and Solution of an Infinite Phased Array of Slots Fed by Coplanar Waveguide Over a Dielectric Half-Space 760-6MG-092	Dr. Donald Hanson (1986)
43	A New Measure of Maintainability/ Reliability and Its Estimation 760-6MG-019	Dr. John M. Jobe (1986)
44	Report Not Available at this Time 760-7MG-050	Dr. Louis Johnson
45	Signed-Digit Number System for Optical Adaptive Processing 760-7MG-015	Dr. Panapkkam Ramamoorthy

58	Signal Processing in EW Environment 760-6MG-135	Dr. John Y. Cheung (1986)
59	Report Not Available at this Time 760-7MG-081	Prof. William K. Curry
60	Implementation of Blackbroad Systems in Ada 760-7MG-010	Dr. Verlynda S. Dobbs
61	Surface States and Electron Transport Properties in Semi-Insulating Gallium Arsenide 760-7MG-049	Dr. Narayan C. Halder
62	Investigate Feasibility of Inplementing Associative Memories Using Luminescent Rebroadcasting Devices 760-7MG-029	Dr. Alastair McAulay
63	Automated Translation of Digital Logic Equations into Optimized VHDL Code 760-6MG-055	Dr. George Zobrist (1986)
Flight Dynamics Laboratory		
64	Analytical Model and Computer Program of F-16 Nose Gear and F-16 ALGS 760-6MG-006	Dr. George Doyle (1986)
65	Report Not Available at this Time 760-7MG-124	Mr. Thomas Enneking (GSRP)
66	Report Not Available at this Time 760-7MG-115	Dr. Oliver McGee
67	Development of a Technique for Prediction of Internal Heat Transfer in Actively Cooled Structures 760-6MG-079	Dr. V. Dakshina Murty (1986)
68	Radiation Hypersonic Aerodynamics 760-7MG-121	Dr. Shiva Singh
69	Report Not Available at this Time 760-7MG-088	Dr. Gary Slater
70	Report Not Available at this Time 760-7MG-080	Dr. Forrest Thomas

71	Report Not Available at this Time 760-7MG-102	Dr. William Wolfe
72	A Chemical Kinetics Model for Mach 5 - 14 Hypersonic Flow 760-6MG-109	Dr. Tsun-wai G. Yip (1986)
Logistics Command		
73	Development of a Microcomputer Lateral Resupply Simulation System 760-7MG-116	Dr. Howard Weiss
Materials Laboratory		
74	Development of Expert System Control of a Carbon Fiber Production Process 760-7MG-027	Dr. Bruce A. DeVantier
75	Influence of Microstructural Variations on the Thermomechanical Processing in Dynamic Material Modeling of Titanium Aluminides 760-7MG-077	Dr. Ravinder Diwan
76	Report Not Available at this Time 760-7MG-097	Dr. Bruce A. Craver
77	Report Not Available at this Time 760-7MG-013	Dr. John W. Gilmer
78	Report Not Available at this Time 760-7MG-075	Dr. Gordon Johnson
79	Studies on the Compatibility of Potential Matrix and Reinforcement Materials in Ceramic Composites for High-Temperature, Aerospace Applications 760-6MG-121	Dr. Gopal Mehrotra (1986)
80	Synthesis of Compounds Capable of Intramolecular Cyclization - Aromatization Reactions 760-6MG-065	Dr. Robert Patsiga (1986)
81	Leaky Rayleigh and Lamb Waves on Composites 760-6MG-007	Dr. Nisar Shaikh (1986)
82	Performance Improvement in Knowledge-Based Process Control Systems 760-7MG-044	Mr. John Usher (GSRP)

94	Comparison of Supervisor's and Incumbent's Estimates of SDy 760-7MG-009	Dr. Michael Matthews
95	Report Not Available at this Time 760-6MG-136	Dr. Jorge Mendoza (1986)
96	The Role of Fourier Descriptions for Shape in Visual Form Perception 760-7MG-082	Dr. John Uhlarik
97	Comprehensibility of Technical Text 760-6MG-080	Dr. Doris Walker-Dalhouse (1986)
98	Report Not Available at this Time 760-7MG-046	Dr. Charles Wells
99	Mechanisms of Contrast and Lightness Constancy 760-6MG-051	Dr. Billy Wooten (1986)

Occupational and Environment Health Laboratory

100	Phytotoxicity of Soil Residues of JP-4 Aviation Fuel 760-7MG-059	Dr. Richard H. Brown
101	An Impact Study for the Contracting Out of In-House Analytical Services at the USAF Occupational & Environmental Health Laboratory - Brooks AFB, San Antonio, Texas 760-7MG-096	Dr. Elvis E. Deal
102	Effects of Metal Mutagens on the Synthesis and Accumulation of Macromolecules 760-7MG-001	Dr. Kiah Edwards
103	Development of a Rapid and Sensitive Assay Procedure for the Detection of the Protozoan Parasite Giardia Lamblia in Drinking Water Supplies 760-6MG-062	Dr. Ralph J. Rascati (1986)

School of Aerospace Medicine

104	Limitations to Heavy Work of Personnel Wearing at 21°C: U.S. Military Chemical Defense Ensemble 760-7MG-067	Prof. Phillip A. Bishop

105	Report Not Available at this Time 760-6MG-118	Dr. Hoffman Chen (1986)
106	Nucleic Acid Hybridzation - Dot Blot Test for the Presence of Ureaplasma Urealyticum and Mycoplasma Hominis 760-6MG-076	Dr. Vito DelVecchio (1986)
107	Report Not Available at this Time 760-7MG-078	Dr. Mohammed Maleque
108	The Asymptotic Description of Precursor Fields in a Causally Dispersive Medium 760-7MG-033	Dr. Kurt Oughstun
109	Report Not Publishable at this Time 760-7MG-091	Dr. Ralph Peters
110	Model Systems for Assessing the Effects of Microwave Radiation on the Immune System 760-7MG-060	Dr. Stephen Pruett
111	Report Not Available at this Time 760-7MG-043	Dr. Wesley Tanaka

FINAL REPORT NUMBER 26
REPORT NOT AVAILABLE AT THIS TIME
Dr. Charles Bump
760-7MG-076

A Report to

U.S. Air Force Office of Scientific Research

THE OMNIDIRECTIONAL TORQUER - EXPERIMENTAL PROTOTYPE MODEL I

A Research Initiation Grant funded for the calendar year 1988

Principal Investigator: Stephen J. Gold, PhD.

Department: Electrical Engineering

Institution: South Dakota State University; Brookings, South Dakota

Technical Focal Point: Major Steven Lamberson, PhD
Director, Large Space Structure Program
F.J. Seiler Research Laboratory
U. S. Air Force Academy
Colorado Springs, Colorado

Date Submitted: 29 December 1988

Signed: _____

Stephen J. Gold, PhD.
Principal Investigator

2 Copies to: Universal Energy Systems, Administrators for
U. S. Air Force Office of Scientific Research

1 Copy each for: Electrical Engineering Department, SDSU

Engineering Dean's Ofiice, SDSU

Center for Powere System Studies, SDSU

Engineering Experiment Station, SDSU

Grants Office, SDSU

2 Copies for the Pricipal Investigator

Summary

This is a report of work done in the Electrical Engineering Department at South Dakota State University throughout the calendar year 1988 on the Omnidirectional Torquer Project. A research initiation grant was secured from A.F.O.S.R. in December, 1987, under which the principal investigator, Stephen J. Gold, PhD, Associate Professor of Electrical Engineering at SDSU was encouraged to proceed with his plans to build a working prototype of this 3-degree of freedom induction machine. It has a spherical rotor, and its useful output is the countertorque the rotor exerts on the stator when the rotor is accelerated. Preliminary design studies for the concept were done during the summer of 1987, when the principal investigator had a summer faculty research position at the F. J. Seiler Laboratory in the U.S. Air Force Academy, Colorado Springs. The report of August 1987 was entitled "Design of an Omnidirectional Torquer".

Notification was received that this second stage of the project would be funded on 20 December 1987. The sole-source vendor for the most expensive and critical parts was immediately contacted. Those parts were custom-machined hollow ferrite hemispheres for the rotor core. After some preliminary negotiations, they were formally ordered in early February. Initial contacts with Ceramic Magnetics Co. of Fairfield N.J. indicated that they could deliver such custom made part 12 weeks A.R.O.; but they quoted 16 weeks which would have got them to us by late May. They finally arrived on 23 July! Since most of the work was supposed to take place during the summer, this late delivery put the project pretty far behind schedule.

In this report is a brief review of the history of the idea and uses for a 3-D motor. This is followed by a section about smooth-rotor machines, which is a different design concept than the one originally explored in the first design report. Then there's a short section discussing the electronic control system for this machine. Next are two pages of photographs and a narrative describing the construction of Experimental Model I. The report ends with a description of the performance test that was done on Model I on 23 December 1988. Unfortunately, it did not work. The reason is thought to be binding between the rotor and the stator caused by distortion of the slightly-flexible stator as it was hung suspended for the test. We will now modify the design using the same rotor in an Experimental Model II which will be built this spring. The new design will feature a stiffer stator with more clearance.

History of the Spherical Rotor Machine Idea

The concept of having a reaction sphere instead of three separate reaction wheels for spacecraft attitude control was first suggested in the early 1960's. The history of these early proposals is summarized in the first four pages of Ref. (1), which Major Lamberson and others a F. J. Seiler Laboratory have, so they will not be discussed further here.

Present interest in the concept comes mainly from a need to kill the vibrations that arise during slewing maneauvers of large satellites. These satellites may consist of several modules connected by slender structural members. Because the structural members have small cross-sectional area, but do tie together large masses, they cannot exert a lot of force so the masses must be accelerated rather slowly. Whenever such a satellite is re-oriented in space, its natural vibration modes will be excited. Those vibrations tend to be slow (in the sub-audio range, perhaps as slow as 0.1 Hz) because of the slender members' inability to rapidly accelerate large masses. If the structural members are elastic, there is almost no natural damping, and the satellite would vibrate for a long time. This calls for an active vibration damping mechanism.

The scheme we envision would use torque to unbend a flexing member in a timed way that would quickly quell vibrations. Using torque is more effective than using force for this, because torque is a vector concept (actually a pair of equal-but opposite force vectors separated by a moment-arm) that can be fairly freely translated from one place to another in a spacecraft; but a force needs to be applied at a specific point. Furthermore, unless the force passes through the center of gravity, it will also have the effects of torque. On the other hand, it is not necessary that a torquer-motor be located at or even near the center of gravity for it to work.

Torque for vibration damping would be in the form of '+' for a couple of seconds, then '-' for about the same length of time, then'+' again. The size of torque needed would diminish so that the magnitude of each peak would be less than the previous peak (damped sinusoids). On the other hand, inertial attitude controls on a satellite spin up a flywheel in one direction and hold that same direction of rotation for as long as it takes for the conservation-of-angular-momentum principle to roll the surrounding spacecraft around to its new desired orientation. With both attitude control and vibration damping, it's the countertorque that the rotor exerts on the stator while the rotor is

being accelerated that does the job. General vibration damping and attitude control can be accomplished by using three inertia wheels with independent axes, each driven by its own torquer-motor. If more than one inertia wheel turn at the same time, there will be some gyroscopic interaction effects that require sophisticated control coordination. With a spherical rotor reaction machine, only one torquer is need for the whole spacecraft, and gyroscopic cross-coupling effects are much less. The mass of a single spherical machine would probably be less than three cylindrical machines, too.

This writer has no knowledge of anyone who has actually built a spherical rotor induction motor, although the idea was mentioned in the literature of the early '60's. The attempt through the calendar year 1988 to produce a working prototype exposed the principal investigator to some of the practical difficulties associated with making such a device: The problem of how to keep the rotor centered seems easy enough; all that is necessary is to suspend the rotor ball with neutral bouyancy in a liquid, start it spinning, and hydrodynamic forces will pull it toward the center. The more important practical problem is how to make/keep the rotor cavity spherical in a hand-made one-of-a-kind prototype that was assembled using thin & flexible parts. The practical 'plumbing' problem of getting the fluid to/from the gap surrounding the rotor (making seals, finding and repairing leaks, routing the wiring around the fluid tubes) was more difficult than anticipated. Getting the required electromagnetic properties of low permeability so the magnetic fields would be easily established, and high resistivity so eddy-currents would be minimised--combined with structural strength, was tough. Finally, overcoming gravity; not only as it acts on the rotor, but acting on the whole thing and deforming it somewhat as it hangs or sits, was an unforeseen problem.

We still believe it can be done, but we now know it won't be as easy as the first theoretical study indicated.

Theory of Operation for Smooth Spherical Rotor Induction Machines

This part reports investigations undertaken in the spring of 1988. Here we explain in simple terms familiar to anyone who has studied cylindrical rotor induction motors, how a spherical rotor machine works. At the end of the section is a page that explains how the magnitude and direction of the reaction torques can be controlled.

In the usual cylindrical rotor a-c machine, the spatial distribution of magnetic flux around the rotor periphery should be sinusoidal, i.e:

$$B(\theta) = B_m \sin(N\theta)$$

where B_m is the maximum flux density, in $^{wb}/_{m^2}$
θ is the spatial angle, measured from a reference point on the rotor, and N is the number of North magnetic poles in the machine. Some cylindrical rotor motors have 2, 3, 4 or more north poles, but Fig. 1 shows only one North pole since that is how many there are on the spherical rotor machine discussed below. The flux pattern of the spherical rotor machine is similar to that ffor a cylindrical rotor. In fact, in a diametrical cross section through the poles and the center of the sphere, its the same picture--see Fig 1 above. For a machine with a spherical rotor, the ideal flux distribution is described mathematically by

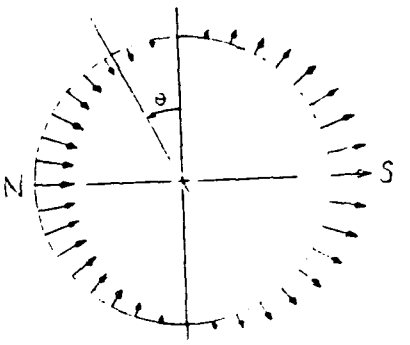


Fig. 1 Flux Distribution for Both Cylindrical and Spherical Rotor A-C Induction Motors. Diametrical Cross Section is Shown.

$$B(\theta) = B_m \sin(\theta) \qquad Eq(1)$$

where θ is the angle measured from the 'equator' great circle plane midway between the maximum flux density magnetic poles, and a ray to the point in question drawn from the center of the sphere. In the language of those who draw maps on spheres, the ideal flux density for a spherical rotor is zero at the equator, and changes as the sine of the latitude angle, finally reaching $+B_m$ at the noth pole, and $-B_m$ at the south pole.

With cylindrical rotor machines, it is a well-established principle that any departure from the ideal spatial distribution can be analysed using Fouriers series analysis to find the 1st, 2nd, 3rd, etc 'Harmonics' which when added together yield the actual flux pattern. All harmonics other than the 1st have detrimental effects that tend to brake a cylindrical motor. For spherical geometry, there are some analogous basis functions like the Fourier series sinusoids. They are called spherical surface zonal harmonics, which can be faound is most books of mathematical tables. The spherical surface zonal harmonic of order 1 correspond to the fourier fundamental or 1st harmonic; Fig. 1 shows and Eq (1) describes a spherical surface zonal harmonic of order 1. Higher order zonal harmonics have different shapes; for example, the 2nd-order one has a peak at its poles and a valley at the equator, and it can be oriented so its poles are not at the same place as the 1st harmonic's

poles. As with the cylindrical machine, any distribution of magnetic flux over the spherical surface can be described mathematically as a weighted sum of the various spherical surface zonal harmonics, at various angular orientations. The angular orientations are sort of like the phase angles in fourier series, except that there are two angles (azimuth and heading) instead of the single fourier angle. Like the cylindrical case, those parts of the spherical flus distribution that cannot be described by the 1st order spherical surface zonal harmonic are detremental to the performance of the machine.

Finally, it should be noted that both the ideal cylindrical and the ideal spherical flux distributions can be thought of as the radial component of flux density resulting when a non-magnetic sphere or cylinder is in a uniform magnetic field (the cylinder's axis to be perpendicular to the uniform field).

The ideal flux distribution for a cylindrical-rotor motor is closely approximated by using a distributed winding with wires embedded in stator slots. Spherical rotor machines' ideal flux can also be set up using a distributed winding. If it is assumed that essentially all the reluctance of the magnetic flux path is from the gap which separates the rotor from the stator--which is not a bad assumption even for tight-fitted squirrel-cage induction cylinders, and a very good assumption for the long-gap spherical machine discussed herein--then the distributed winding pattern which gives the ideal flux pattern will be turns that are as close together as possible near the equator (wires touching), and the spacing between adjacent turns will gradually increase as the sine of the latitude as one moves toward the poles. Fig. in Ref (1) illustrates such a turns layout. Another way of looking at the ideal turns layout is to consider that there are equally-spaced 'slices' perpendicular to the polar axis, and each slice goes through the center of one turn. Fig. 2 a twenty-turn winding with this property emphasized.

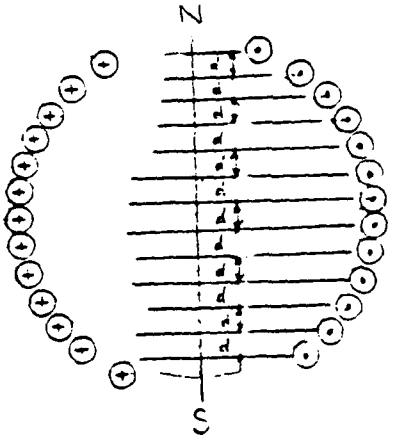


Fig. 2 A Distributed Electromagnet Winding that Will Establish an Ideal Flux Distribution.

Rotating magnetic fields for cylindrical motors are set up by polyphase 2 Ø or 3 Ø currents in two or three separate stator windings. The same idea can be employed with spherical motors. If there are two concentric stator windings whose polar axes are perpendicular to each other; then exciting one

with a current of the form $I_a(t) = I_m \cos(\omega t)$ and the other with a current of the form $I_b(t) = I_m \sin(\omega t)$ will create a rotating ideal flux pattern. To show this we need to develop a concept for spheres which is analogous to spatial phasor of cylindrical machine theory. In cylindrical machines a flux pattern can be represented by a vector of magnitude B_m passing radially through the point of maximum flux density; it is understood that this single vector represents the entire distributed flux pattern. Furthermore, if there are two flux patterns each of which is sinusoidally distributed in space but created by a separate sinding, the combined effect of the simultaneous application of both stator currents can be obtained by doing vector addition of their individual vector representors. If one of these vectors is vertical and varies as $B_m \sin(\omega t)$ while the other is horizontal and varies as $B_m \cos(\omega t)$, the vector sum will have constant magnitude and rotate at synchronous speed. The same idea carries over into spherical geometry: a flux pattern set up by stator current that is a spherical surface zonal harmonic of order 1 can be represented by a single vector passing through its pole, of magnitude B_m. A sign convention similar to the one used for cylindrical motors is used so that radially inward flux is consided to be positive. If there are two such patterns caused by two separate stator windings carrying currents; then the effect of each individual stator current can be shown as a single vector that passes through the peak of the distribution it causes; and the combined effect of both currents acting simultaneously is found by vector addition of their separate individual vectors. This very important concept can be simply derived as is shown below.

Let the flux pattern vectors be $\vec{A} = (A_x \quad A_y \quad A_z)$ and the 3-dimensional rectangular cartesian vector $\vec{B} = (B_x \quad B_y \quad B_z)$ be the other. The radial component of the flux distribution A that passes through a general point on the spherical surface can be evaluated by doing a dot-product operation multiplying the direction cosines of $\vec{G}$ times the components of $\vec{A}$. If the direction cosines of $\vec{G}$ are (d e f), that dot product operation yields the scalar result $d A_x + e A_y + f A_z$. Similarly, the radial flux at G attributable to the flux distribution of B can be expressed $d B_x + e B_y + f B_z$; and the combined effect is $(A_x + B_x)d + (A_y + B_y)e + (A_z + B_z)f$, which is the same result we obtain if we do the vector summation first, and then the dot product. The idea can be extended to the combined effect of two stator currents and one rotor current pattern. The additive

property implies that a pair of stator windings surrounding perpendicular axes excited by quadrature currents of the form $I_a(t) = I_m\cos(\omega t)$ and $I_b(t) = I_m\sin(\omega t)$ will by their joint action create a rotating flux pattern whose magnitude doesn't change but whose pole moves around at ω $^{rad}/_{sec}$. The pattern of the rotating flux is the ideal spherical surface zonal harmonic of order 1. It rotates around the third axis which is found by doing a vector cross product of the two coils' axes.

Consider next the pattern of currents that is induced on the rotor surface by the above-described rotating magnetic field. The rotor has a highly permeable ferromagnetic core whose electrical conductivity is negligable, and is is clad on the outside by a metal layer (silver) with high electrical conductivity and permeability about the same as air. Suppose there is relative motion between the rotor and stator magnetic field described by mechanical angular velocity vectors $\vec{\omega}_{rel} = \vec{\omega}_s - \vec{\omega}_r$

The relative motion will induce electromotive forces (voltage) in the metal on the surface of the rotor, and these voltages will cause currents to flow. The voltage gradient can be computed from

$$\vec{E} = \vec{U} \times \vec{B} \ ^{volts}/_{meter} \qquad Eq(2)$$

which can be integrated along a path to get volts of motional EMF. $\vec{E}$ is always tangential to the surface if $\vec{B}$ is radial. The pattern of the induced $\vec{E}$ field and also of the induced currents is shown below in Fig. 3. The pattern of the induced currents is the same as the stator current pattern: the current in each of the hoops sliced out by equally-spaced planes normal to the rotor axis are equal. This fact was uncovered in the investigation by indirect means. The induced voltages in the rotor surface conductor were computed numerically by doing point-by-point vector cross products per Eq (2) on a computer. These values were then integrated numerically using Bode s five-point integration formula to find the voltage induced in a length of hoop. The voltage was divided by the resistance of that length of conductor to find a current flowing there. Currents were also numerically integrated across the

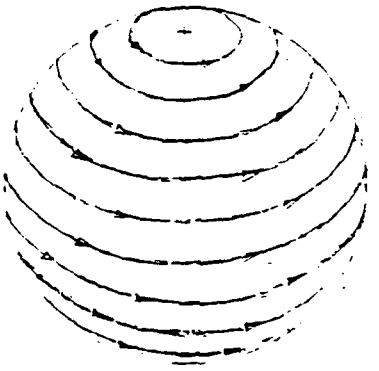


Fig. 3. Rotor Currents Induced by Relative Motion with Respect to an Ideal Flux Pattern Set Up by a Quadrature Pair of Stator Currents.

width of a strip to get a weighted average current-flow. When all of this was put together in a common program, it appeared that the current in each of the hoops was the same. When more and narrower hoops were used, and more lengths of integrated voltage, the agreement got better and better until with 20 hoops in each hemisphere all hoop currents were the same to within 5+ significant figures for all hoops except the last one whose one boundary was the pole-point. When the number of hoops and the number of integration lengths were both doubled, the agreement between computed hoop currents improved by nearly two more digits, as would be expected with a 6th order numerical integration algorithm like Bode's five-point method as it converges to the ultimate solution. With the pattern of induced rotor currents being the same as the stator currents' pattern, it follows that the magnetic field set up by the rotor currents will also be a spherical surface zonal harmonic of order 1 which can be represented by a single vector. That vector moves with respect to the already-moving rotor at a velocity such that it is going at synchronous speed relative to the stator, and in fact follows the stator field around just like in a cylindrical motor. Since stator and rotor fields are both turning at the same speed around the same axis, the vectors representing them can be added together to get a composite vector caused by two stator windings and the rotor; this composite vector quantifies the gap flux density which determines the degree of saturation in the ferromagnetic parts of the machine.

Therefore we see that a spherical machine can be understood by extending the ideas commonly used to explain cylindrical motors. The spatial distributions are different, with spherical surface zonal harmonics of order 1 for the flux and equal hoop current causing that flux, instead of the sinusoidal flus and current spatial distribution of the cylindical case; but the time-variation of currents to create a rotating magnetic field is sinusoidal for both the cylinder and the sphere. If the sphere has three concentric windings, the stator field can be made to rotate around either x- or y- or z- axis, whereas the cylindrical motor has only one degree of freedom and its magnetic field must rotate about the motor shaft.

We have developed a method that can generate a magnetic field which rotates about any axis with arbitrary orientation in 3-dimensional space, by using properly sized and timed currents simultaneously in all three of the concentric stator windings. The idea is developed below:

Let (a b c) be the direction cosines of an arbitrarily-specified axis of rotation. Since angular velocity is a vector, a rotating field that

goes around the arbirary axis can be synthesized by adding together properly-sized magnetic fields that to go around the x-,y-, and z- axes. To get a magnetic field to go around the x-axis requires in the windings that encircle the y- and z- axes. Positive (counterclockwise) rotation about the x-axis of a right-handed cartesian frame could be set up by

$$I_x = 0 \qquad I_y = a\ I_m \cos(\omega t) \qquad I_z = a\ I_m \sin(\omega t)$$

Similarly, a pair of currents that would cause positive rotation about y-axis is $I_x = b\ I_m \sin(\omega t)$ $I_y = 0$ $I_z = b\ I_m \cos(\omega t)$

Finally, the required currents for the z-component of the rotating field are

$$I_x = c\ I_m \cos(\omega t) \qquad I_y = c\ I_m \sin(\omega t) \qquad I_z = 0.0$$

Where I_m is the size of the current that would be needed to create the requisite magnitude of rotating field if only two windings were used to do it. Note that since a, b , and c are direction cosines, $a^2 + b^2 + c^2 = 1.0$ The currents required in each winding can be expressed in terms of sines and cosines as

$$I_x = I_m\ c \cos(\omega t) + b \sin(\omega t)$$

$$I_y = I_m\ a \cos(\omega t) + c \sin(\omega t)$$

$$I_z = I_m\ a \sin(\omega t) + b \cos(\omega t)$$

or in terms of magnitudes and phases as

$$I_x = (b^2 + c^2)^{\frac{1}{2}}\ I_m \cos(\omega t - \alpha) \qquad \alpha = \mathrm{Tan}^{-1}(b/c)$$

$$I_y = (a^2 + c^2)^{\frac{1}{2}}\ I_m \cos(\omega t - \beta) \qquad \beta = \mathrm{Tan}^{-1}(c/a)$$

$$I_z = (a^2 + b^2)^{\frac{1}{2}}\ I_m \cos(\omega t - \gamma) \qquad \gamma = \mathrm{Tan}^{-1}(a/b)$$

Putting 'sin' in where 'cos' appears in the above expressions has the effect of delaying everything by ¼ cycle, but the orientation and magnitude of the resulting rotating field would be unaffected. It is preferable to use sines rather than cosines because when the argument of a sine function is zero, that corresponds to a zero-crossing instant. The sized/timed sinusoidal currents described above can be generated by solid state controlled current sources whidh use pulse-width-modulated square waves to approximate sine waves. Such sources can be designed to keep pollution from low order harmonics very small.

The smooth-rotor induction sphere works better than a sphere with a surface structure of triangular interlaced grid sectors, such as the one that was analysed in Ref. (1). The induced current in a solid cladding is free to flow along the most direct route, collinear with the induced voltage:

and the Laurentz magnetic forces and torques acting on those currents are directly additive rather than adding in a jagged vector sense. Therefore it is not surprizing that the smooth rotor suface should produce more torque (almost 1½ times as much) than was produced with the triangular-grid-covered rotor that was studied in Ref. (1). For a peak flux density of 1.0 $^{w}/_{m}2$, the triangular grid rotor was computed to produce 2.1 n-m of torque with 400 Hz excitation, but the smooth rotor will yield nearly 3.0 n-m for the same conditions, according to computer simulations done in April, 1988.

Electronics for Monitoring and Control of the Omnidirectional Torquer

Some students have become affiliated with the Omnidirectional Torquer project. Their interest is mainly in the measuring and control electronics and in the algorithms that are embodied in the microprocessor program, that are needed to make a closed loop control system with the Torquer as the central element. The Principal Investigator felt that a lot of this effort was getting the cart before the horse; that the main thing was to get the torquer itself working so its parameters (current requirements, output torque) could be established before delving too specifically into designing an electronic digital feedback system to go around it. Nevertheless, the students needed something to do last spring and summer and fall while the torquer was being procured and built. This section is a brief outline of their efforts.

A. Prabhakar, a graduate student, worked on sytem design. Beginning with the strain gages which measure torque, he designed manually balancable bridges for them and instrumentation amplifiers to get a signal big enough to work with. He started with the assumption that the torquer could cause a strain signal of up to ± 0.1 mv, and designed instrumentation amplifiers to get that up to ± 5.0 volts, as needed for analog to digital conversion. Since there are twelve individual strain gages on the torquer test stand, an analog multiplexer was procured that can be commanded to select the individual channel and read it in about 0.1 msec. All twelve channels are to be read sequentially within a time-frame of 1.5 msec, which for now we are regarding as simultaneous sampling since that is much faster than the expected mechanical motion of the torquer. Analog to Digital conversion is being done with a successive approximation type 10-bit A/D converter. The output of the A/D coverter is interfaced through a parallel port into an 8086 microprocessor. For the microprocessor there is a memory board with 32K of EPROM for program storage and 64K of RAM for data storage. A keyboard is included so we can

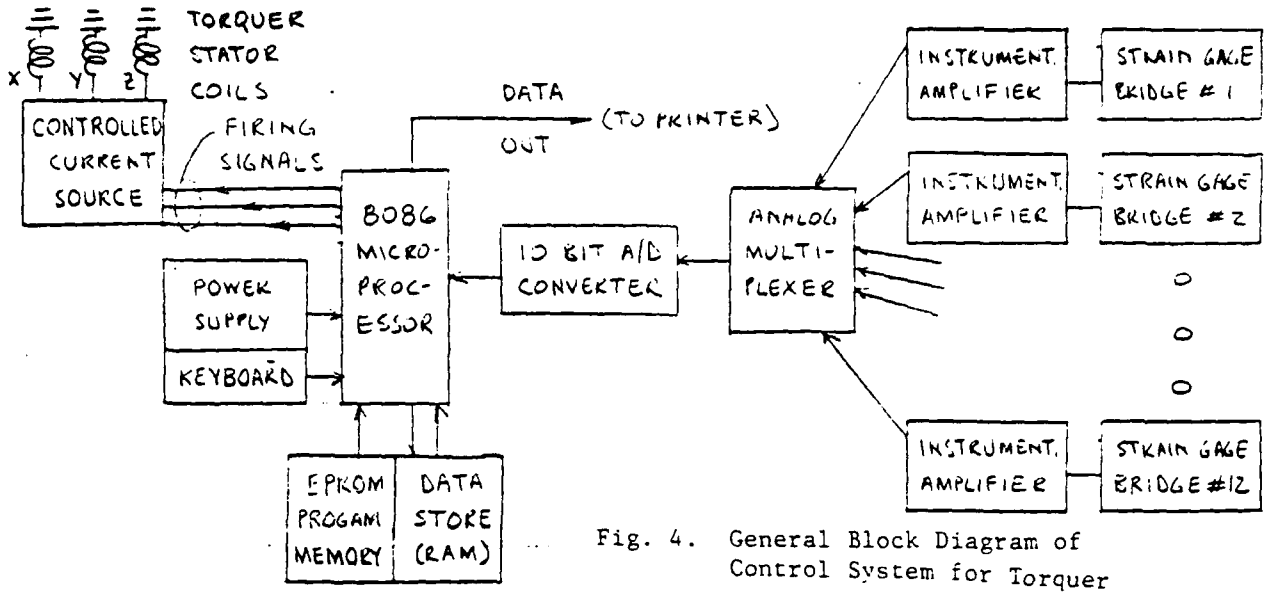


Fig. 4. General Block Diagram of Control System for Torquer

enter the experimentally determined calibration constants that the microprocess will use to convert the measured strains to torques according to

$$\begin{matrix} T_x \\ T_y \\ T_z \end{matrix} = \begin{matrix} a_{11} & a_{12} & \cdots & a_{1,12} \\ a_{21} & \cdot & \cdots & \cdot \\ a_{31} & \cdot & \cdots & a_{3.12} \end{matrix} \begin{matrix} s_1 \\ s_2 \\ \vdots \\ s_{12} \end{matrix}$$

The computed torques will be compared against desired or commanded values, and the smpled and quantized error values E_x, E_y, and E_z will be used to command more or less torque around each of three perpendicular torquer axes as needed to reduce the errors. The digital electronics system has a parallel output port that can be used to move data from its memory bank onto a personal computer disk or to a printer for analysis of performance.

Jacob Kurian is an undergraduate student. He has been working mainly on fabricating the circuit boards for the instrumentation amplifiers and the digital electronics system.

X. Shen is a graduate student who wanted a design project to work on. We have not paid him any money for his efforts, but he has designed a constant current source solid state motor controller that works on the principle of pulse width modulation to provide up to 10 $Amps_{rms}$ from a $\pm$ 100 V_{dc} supply which can be used for exciting the x-, y-, and z- windings of the torquer stator. His design employs power MOSFET transistors.

H. Chau will be a new graduate student who starts in January, 1989. We have approached the E.E. Dept. Head and the Director of the Center for Power System Studies to see if we can supply him with a small research

Photo 1. One of the Hollow Ferrite Hemispheres that are in the core of the rotor (four were furnished, two were used.) These parts did not arrive till 23 July.

Photo 2. The 1mm thick plastic (polyethelene) ball that was cut in half, smoothed inside, and used for the stator inner sheath.

Photo 3. Silver-clad rotor sphere fitting inside half of the stator inner sheath.

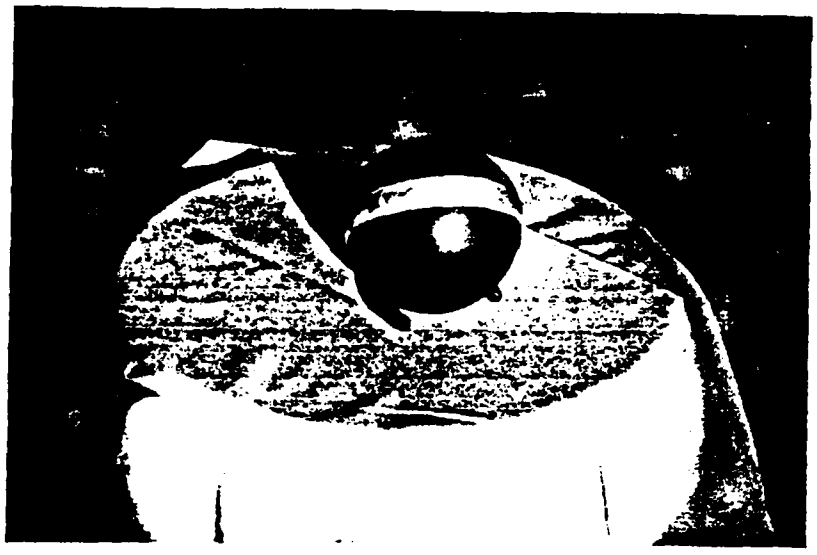


Photo 4. Stator sheath closed with Equatorial Seal. Silver clad rotor ball inside can be seen through open tube-port.

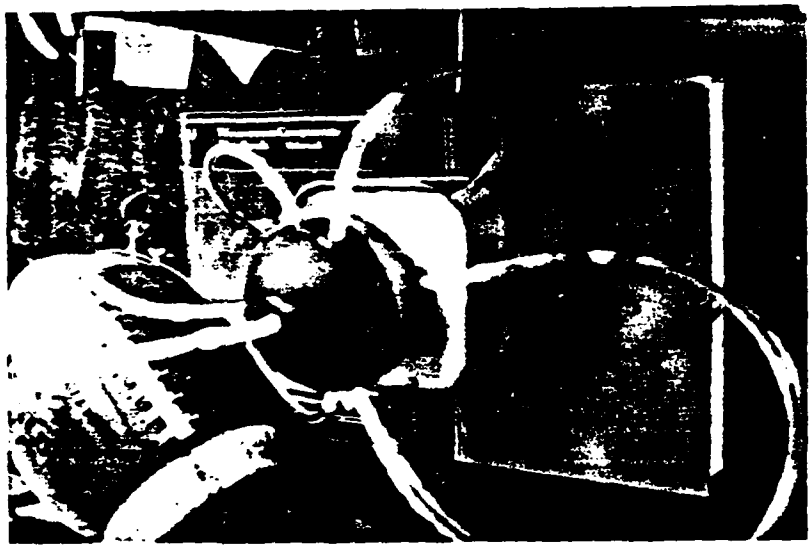


Photo 5. Stator sheath with cooling fluid tubes attached to the eight ports, being subjected to a pneumatic leak test. Football serves as pressure reservoir.

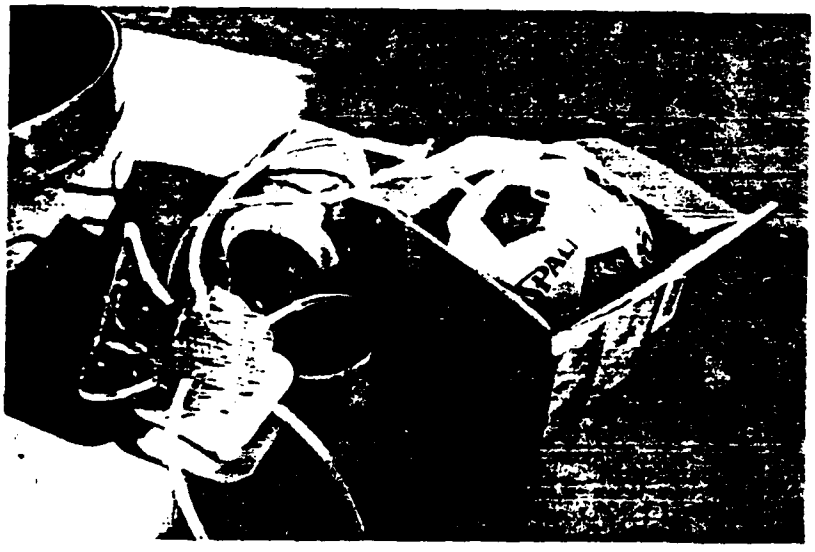


Photo 6. Assembling the stator windings while the inner sphere was pressurized. Soccer ball and football were both used so one could be removed and pumped up while the other held pressure.

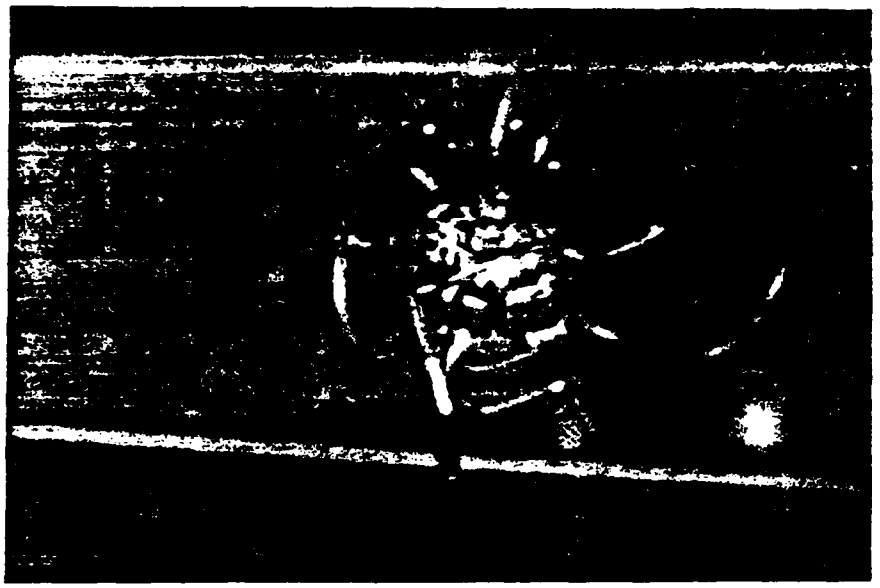


Photo 7. The torquer with some of the stator yoke ribbon in place. This was its state of construction when it was subjected to the first performance test on 23 December, 1988.

PHOTOS 1 - 7 SHOW SOME STAGES IN THE CONSTRUCTION OF THE TORQUER

stipend for spring semester, 1989. He will work to bring the electronics together and interface it with Experimental Model II of the torquer, which will be built this spring.

The preliminary performance testing of the omnidirectional torquer does not require any electronics. We have 60 Hz 3 Ø power available from the local utility, and a 400 Hz 3 Ø generator is also available. 3 Ø power can be passed through a 3 Ø autotransformer, and then through a scott-tee transformer bank to make variable magnitude 2 Ø power at either 60 or 400 Hz. Applying this to any two of the three stator windings should make the torquer go.

Construction of the Omnidirectional Torquer (August - November, 1988)

The Interim Report, sent to Major Lamberson the first week of August, stated that the ferrite hemispheres had just been received from the vendor (Ceramic Magnetics; Fairfield, New Jersey) and that their delivery was about seven weeks later than promised. Upon receiving them, we sanded smooth/flat the mating surfaces and epoxied the hollow halves together. Then we checked the outside for the requisite sphericity with a micrometer, sanding with fine emory paper where necessary to obtain the diameter of 9.60 cm. Finally it was hand sanded smooth to a 400-grit finish, leak-checked by immersing it in water (no bubbles were seen), and then sent by priority mail to the Jay Electroforming Company in Chicago. That had agreed to put the 0.6 mm thick silver layer on the outside surface. We chose them because they offerred a one-step process, rather than vacuum depositition followed by electroplating that would have been required by other vendors. Also, they said they could get it back to us 'in a few days'. The 'few days' turned out to be 4+ weeks, and the silver-clad rotor did not arrive back in Brookings until 23 August--only six days before fall semester classes started here. When the clad sphere arrived, its diameter was measured with a micrometer and found to exceed 3.85" (= 9.78 cm) in some directions. In fact, it was too big to fit in the spherical cavity we had made for it. Nearly 50 hrs of hand-sanding with 180- and 400-grit emory paper was required to round/smooth it to diameters that were measured to be between 9.72 and 9.73 cm in all directions. That done, it did fit inside the stator sheath with about 1.0 mm clearance.

The polyethene ball that was used for the stator sheath originally had a wall thickness of 1.0 mm. When whole, this ball was quite rigid and could not be deformed a millimeter without the application of several newtons of force. We cut it half with a razor blade, and smoothed the inside, which had some ridges and indentations using 80, 220, and 400-grit emory paper until it was smooth; sphericity was checked on the whole ball before it was cut in half. These sanding operations removed about 1/3 of the mass of the stator inner sheath: when it was whole, it weighed about 37 grams; after the smoothing operations were done on them, each half weighed only about 13 grams. Eight holes were made in the stator sheath to be inlet/exit ports (four of each) to distribute $ZnBr_2$ solution evenly around the gap. Eight ports were decided on instead of only two because of concern that with a single inlet and outlet, flow pressure differentials would push the rotor ball to one side, perhaps

even blocking the exit port. The holes were originally fitted with short (5 cm) glass tubing pieces which were sanded on the inside to be smooth and flush with the spherical surface there, and sealed with a bead of epoxy around the outside. These glass tubes were replaced in the second building of the stator with semi-rigid polyethelete tubing (see below) because one of them had been chipped at its outer lip.

The rotor ball was put inside the stator sheath, and the equatorial seam was resealed from the outside with a 2 cm wide strip of epoxy. When the epoxy was partially set, but still tacky, it was reinforced with a single layer of fibre tape to give the equatorial seam about the same strength and rigidity the hollow ball had before it was cut open. The stator sheath with its protruding inlet/outlet tubes was allowed to set up overnight before subjecting it to a 10 p.s.i. pneumatic leak test. The eight tubes were sealed by joing one to the other with flexible tubing; except for two. One of them was fitted with a piece of flexible tubing which was triple-kinked to make a seal, and the other was connected via a 3' lenght of flexible tubing to a soccer ball which served as the 10p.s.i. pneumatic reservoir. The stator sheath was then immersed in a tub of water. Two small streams of air bubbles were obseverved to be coming out near the bases of two of the tubes, and these leaks had to be repaired; but there were no leaks at the equatorial seal where the two spherical halves were joined. At this point in the construction, the ball-within-a-ball could be subjected to sudden twist-jerks by quick hand-wrist action, and it was easily determined that the outer sheath was free to move with respect to the ineer rotor-ball.

After the stator sheath was sealed, the next step was to wrap the A.W.G. #16 copper magnet wire coils around it. Experience gained during the winding of the first coil indicated two things: First, the eight protruding tubes necessitated a significant digression from the simple equal-axial-spacing of the winding-turns that the theoretical zonal harmonic design called for. The original design had 37 turns covering each hemisphere, and a total of 74 turns comprising each of the x-, y-, and z- layers. As built, there were only 35 turns per hemisphere. The egress tubes are spaced equally over the surface of the sphere, like the eight corners of a cube; and to avoid them it was necessary to omit two of the turns about 35^{o} above/below the plane of maximal diameter for each of the three concetric windings. Also, some of the turns had to be a little closer together than the theoretical design called for.

Second, because of the stiffness/springiness of 16-gauge magnet wire, it was found to be very difficult to wrap coils of wire on a sphere while progressing from equator to pole; the wire simply wouldn't stay in place long enough for the epoxy to set. On the other hand, it was relatively easy to put the wire turns in place while progressing from pole to equator because there was already a turn in place on a smaller diameter to help keep the most-recent turn from moving. Because of the difficulty in going one way, compared to the ease of going the other, a construction technique was adopted that laid the windings down as hemispherical half-sindings, with four wire-ends brought out for later electrical interconnection, instead of only two wire-ends. The first time the stator was assembled, all three (x-, y-, and z-) widings were laid down turn by turn and each turn was stuck in place with 5-minute epoxy. One had to wait for the epoxy to set before the next turn could belaid down, which mad this a rather lengthy process. It took more than eight hours for each half winding; about 50 hours in all, to put them firmly in place. The second time this operation was done, Super Glue was used,which sets up in about 10 seconds under the right conditions, so doing the windings took only about 30 hours.

When all the windings were on, the next step was to lay on the vanadium permendur magnetic tape to make the stator yoke. The original concept was to use machined ferrite for the stator yoke, but our single-source vendor (Ceramic Magnetics Co.; Fairfield, N.J.) said it would cost almost $5000. to make stator yoke pieces, which would have exceeded the budget by an uncomfortable amount. So an alternate design using magnetic tape (which cost only about $250.) was adopted. As Supplied by the vendor, the bare magnetic tape was .002" thick, ½" wide, and 1750' long. In order to provide an insulation layer against eddy currents and to aid in sticking the tape to the ball, the metal tape was stuck near one edge of a .375" wide strip of mylar 'Magic Mending' tape from 3M Co. The .125" of sticky surface held the tape to the layers below. These combined tape strips were prepared in strips of 20-30 length and wrapped around the outside of the stator windings like yarn around a baseball. We were well into this process (300+ feet of tape laid down) when we began to notice that the torquer unit could no longer pass the aforementioned Twist-Jerk test. Its mass had grown from 1230 g. to more than 2 kg. and its size from 10cm dia. to more thna 12 cm, so this basic, simple test was not as easy to execute now. But using both hands, it was still possible to do it. By the first week in November, it was reluctantly admitted that the device was failing the Twist-

Jerk test. Apparently the wrapping of layer after layer of stator yoke ribbon had exerted inward pressure sufficient to cause deformation of the stator inner sheath to the point where it was in intimate contact with the rotor in at least two places, and binding it so it could not move with respect to the stator. We tried pressurizing it as a possible way to restore clearance, but to not avail; it still could not pass the twist-jerk test with 10 p.s.i. pressure applied. Since that is the pressure that would be available from the $ZnBr_2$ solution pump, there appeared no hope that it would work as built the first time. So we decided to take it apart and build it anew.

The second time, it was assembled pressurized to about 8 psi while the windings and the stator yoke ribbon were wrapped. The twist-jerk test was done at least daily, and it was still passing that test when it arrived at the point in the construction when about 20% of the stator yoke wrapping was on. A conscious effort to not wrap the stator yoke so tightly was adhered to in the second attempt. After Finals Week, in mid-December, it was decided to try some preliminary testing in an effort to get something positive to report at year's end. By 21 December we thought we had enough of it built to give it a try.

Preliminary Performance Test (21 - 23 December 1988)

Not much work was done on the omnidirectional torquer during Finals Week (December 12 - 19), but finally on 21 December there was enough stator in place to merit giving it a try. We needed some experimental results to send in this report.

Measurements of the windings' resistance and reactance at 60 Hz were made. The coils' resistance was measured using a Kelvin bridge, and found to be close to 0.24 ohms, which is the expected value for 18 meters of AWG #16 copper wire at room temperature. When supplied with 5.4 V_{rms} 60 Hz, it was observed that 4.7 A_{rms} of current flowed, and 9.2 watts of power was consumed in each winding. Since

$$I^2 R_{eff} = 9.2 \text{ Watts}, \quad R_{eff} = 0.42 \text{ ohm of effective a-c resistance.}$$

The difference between 0.42 and 0.24 ohms is attributable to the referred-to-the-stator resistance of the silver on the rotor surface. Because of the materials in the core and the stator yoke, the eddy currents in these parts would be negligable. The testing described above is akin to the blocked rotor test commonly applied to cylindrical rotor machines to determine their equivalent circuit parameters. From the complex power triangle,

$$S^2 = P^2 + Q^2$$

where S is 1 Ø Volt-Amperes and P is power, watts; we can find Q (VARs) to be

$$Q^2 = (4.7A \times 5.4V)^2 - 9.2W^2 \text{ , giving } Q = 23.65 \text{ VARs} = I^2X_L$$

$$\text{hence } X_L = j\omega L = j\,377\,L_L = 23.65/4.7^2 = j\,1.07 \text{ ohms}$$

from which the leakage inductance L_L is computed to be $L_L = 0.028$ H. Note that what was determined here was the combined leakage inductance $j\omega (L_{11} + L'_{12})$ due to flux linking the stator winding but not the rotor + flux linking the rotor only. To get an estimate of the self-inductance of the stator winding, it is necessary to run an induction machine at near synchronous speed so ther are almost no induced voltages/currents in the rotor (slip approaching zero). With a 60 Hz supply, synchonous speed for this machine would be 3600 RPM; and there were never any plans to run it anywhere near that fast because gap fluid friction would be excessive at that speed. The fact that combined leakage reactance was almost 3 x as the effective a-c blocked rotor resistance is significant, because that means that at the beginning of startup there is at least

$$(W_\emptyset)_{peak} = L_L(I^2_{rms}) = 0.063 \text{ Joules/winding}$$

of energy being stored in the magnetic field of each coil. Now, the moment of inertia of the hollow spherical rotor is $J = 1.05 \times 10^{-3}$ Kg-m^2, so the kinetic energy stored in it when it is turning at 10 rad/sec is 0.053 Joules--the same order of magnitude as the energy stored in the magnetic field. We hoped that some of the magnetic field energy would be converted to rotational kinetic energy, and that the rotor would acheive a speed of a few RPM after the machine had been energized for a few seconds by a 2 Ø rotating magnetic field.

On 23 December, 1988, a simple test rig, depicted in Fig. was built. $ZnBr_2$ solution was blended to the density (2.28 g/cm3) of the rotor, and fed through a hydrostatic pressurizing tube into on of the stator inlet ports. The other seven ports were either shunted to another port or crimped tightly shut. A standing column of $ZnBr_2$ solution 3 meters high was established in the vertical tubing section, so the the internal pressure of the gap fluid was about 4.3 p.s.i. above atmospheric pressure. The omnidirectional Torquer itself was hanging suspended from a single string about 1 meter long, and it was about 0.1 m off the floor. Besides the suspension string, the other contacts to the torquer were two lengths of flexible tubing (one to admit the hydrostatic pressure column, and other to bleed off bubbles during the filling process) and four quite-flexible lead wirs bringing current to/from the innermost and the

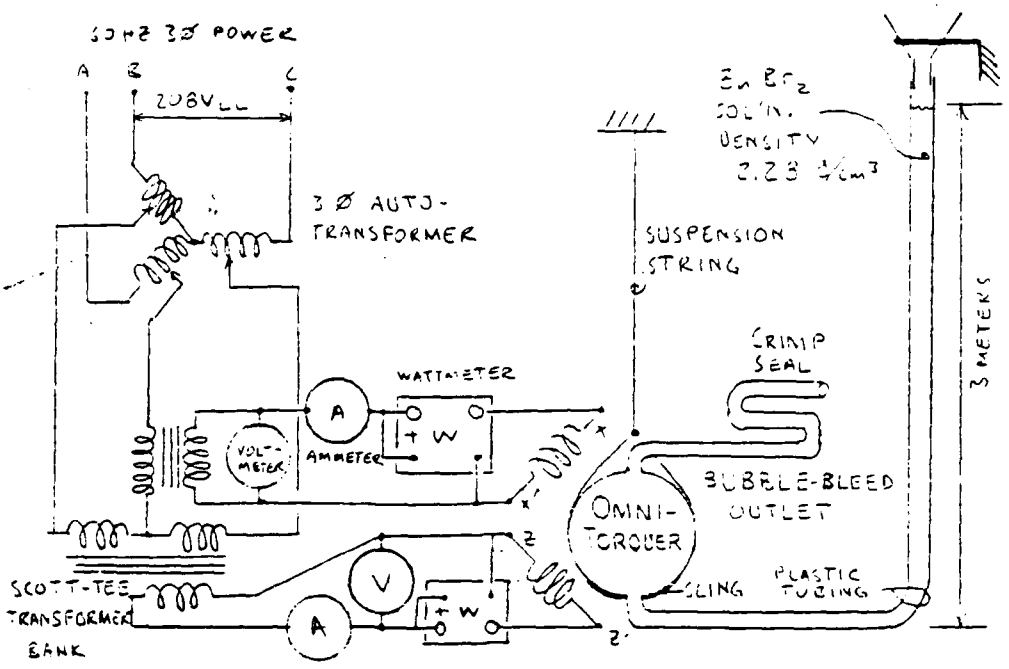


Fig. Hydrostatic Test Rig for Omnidirectional Torquer

outermost of the three windings. The stator windings were supplied with 10 A. of 2 Ø 60 Hz current, and each winding consumed about 36 watts of power. Any countertorque due to rotor acceleration would have disturbed the delicate suspension system in a noticable way.

Unfortunately, when the switches were turned on, no acceleration noises were heard, and nothing moved! In trying to fathom the reason why it didn't work at all; the most plausible explanation is that the suspension sling (which was a double-folded cotton cloth about 4 cm broad) must have created enough distortion in the shape of the stator to cause the rotor to become bound. If this is the reason, then it will be very difficult to come up with a suspension system that won't have similar problems--certainly the tetrahedral test stand we have built would cause probably more distortion at its three bottom contact points. The root of the difficulty is three-fold: 1) The clearance between the rotor and the stator is tight (0.8 mm) even if both parts are perfectly spherical. 2) The completed torquer has a mass of about 3 kg, and we must support that in a 1 G. gravitational field. 3) The stator, built with magnetic ribbon-wraps, is flexible. A way to modify to design that simultaneously attacks root-causes 1) and 3) is discussed in the concluding remarks below.

Concluding Remarks

The theoretical concepts underlying the omnidirectional torquer are fairly straightforward. The induction principle has been used on several other shapes

(induction disks, annular linear induction for liquid metals, two-degree-of-freedom machines) besides the standard cylindrical rotor geometry that N. Tesla patented a century ago. The investigators have full faith that the induction principle can be made to work in spherical geometry, too. The Principal Investigator has brought forth an idea that he would very much like to see come to fruition in the form of a working model. Because the first attempt did not succeed is not sufficient reason to give up the idea. We will try again after modifying the stator design in the following ways: 1) Eliminate the stator inner sheath, which is almost 1 mm thick, to get more clearance between stator and rotor. This can be done by using an intact 10 cm dia. ball as a mold around which hemispherical half-windings are placed and set in epoxy. If the spherical mold is greased first, the epoxy won't stick to it and the ball can be pulled out after the epoxy sets. The hollow hemispheres so created will have wire - for strength - embedded/surrounded by by hard epoxy. If the inner winding layer were fashioned in this manner, it could be a functional replacement for the inner sheath, providing a leakproof pressure vessel with a spherical cavity that is about 3 mm larger than the rotor sphere that must fit inside. 2) Stiffen the stator by encapsulating the whole thing (three concentric windings whose combined thickness is nearly 0.4 cm) in hard epoxy. The wire will act like reinforcing bars in concrete, and the hardened epoxy will be stiff enough so that even a point contact force of 30 newtons from setting the weight of the entire torquer on a hard flat surface will not distort the stator shape enough to cause binding of the rotor inside. 3) To overcome the inlet/outlet differential pressure-push problem while simplifying the laying of windings on the stator, we shall use a single concentric inlet/outlet tube instead of the eight tubes sued in the original design. The gap volume will be nearly tripled, so the volume of fluid will be such that it can absorb a few seconds of rotor heat with no flow at all. When the rotor ball turns, there will be mixing of the newly-injected fluid with that alresady in the gap, and the effluent will be a mixture of cool/heated fluid.

We do not seek any additional funding from the A.F.O.S.R. at this time, but may be getting some graduate student support from local sources (Center for Power System Studies) for spring semester 1989. We will be sending brief Quarterly Updates at the end of March, June, and September, 1989 directly to Major Lamberson, and of course a final report through Universal Energy Systems if/when we have some positive things to report.

References

(1) "Design of an Omnidirectional Torquer", Report of 1987 Summer Research Faculty project by Stephen J. Gold, PhD. Done at F. J. Seiler Laboratory U. S. Air Force Academy,Colorado Springs. August, 1987

(2) "A Critical Comparison of Ferrites With Other Magnetic Materials" Vendor Brochure from Magnetics Division of Spang and Company, (properties of Co-Fe Permendur Tape Cores.) Butler, Pennsylvania 1986

Research Initiation Program Proposal
Final Report

Title: *Calculation of Nonlinear Optical Properties*

Principal Investigator: Henry A. Kurtz

Mailing Address: Department of Chemistry
Memphis State University
Memphis, TN 38152

Proposal Period: Jan. 1 1988 - Dec. 31, 1988

Publications from work:

1. "Calculation of the Nonlinear Optical Properties of Molecules", H. A. Kurtz, J. J. P. Stewart, and K. Deiter, to be submitted to J. Comp. Chem.

2. "Hyperpolarizabilities of Polyenes", H. A. Kurtz, to be submitted to J. Phys. Chem.

Research Summary:

A major goal of this proposal was to test the application of semiempirical based methods for the calculation of the hyperpolarizabilities of polymers. To this end, a series of polyacetylene oligomers was studied, ranging in size from C_2H_4 to $C_{34}H_{36}$. Two conformations for each oligomer was studied: all-trans and all-cis. For the all-trans oligomers the first hyperpolarizability is always zero due to a center of symmetry and for the all-cis forms it is zero for an even number of (C_2H_2) subunits. The second hyperpolarizability results were obtained with a finite-field procedure implemented within the MOPAC program (J. J. P. Stewart, *QCPE Program #455*, 1983; version 3.1 (1986)). The MNDO based results are given in Table 1. For comparison, the results of recent *ab initio* calculations (G. J. B. Hurst, M. Dupuis, and E. Clementi, *J. Chem. Phys.* 89, 385 (1988)) are also given. Results with two different basis sets are shown to demonstrate the spread in computed values.

Of interest is the behavior of the hyperpolarizability as the oligomer length grows. Is there a limiting value of γ/subunit that can be used for polymer calculations? There are two ways to compute a value/subunit of an oligomer properties: 1) divide the value by

the number of subunits (n) or 2) subtract the value for an (n-1) length oligomer from the value for a n subunit oligomer. The former method will be designated /n and the latter /s. The values obtained by these two methods are shown in Table 1. To explore the limiting behavior of these values, $\log(\gamma/s)$ and $\log(\gamma/n)$ are shown in Figure 1. Both these two methods seem to indicate that there is a limiting values but clearly they have not reached it. Following the procedure of the *ab initio* calculations, the $\log(\gamma/\text{subunit})$ can be fit to $a + b/n + c/n^2$. The large n limit is then given by 10^a. This method predicts limiting values of 1.26×10^6 and 1.72×10^6 for the /n and /s methods, respectively.

Another goal of this research was to explore the development of atomic correction factors for the semiempirically computed second hyperpolarizabilities. For the linear polarizability, α, it has been shown that MNDO procedures greatly underestimate the atomic contribution to the polarizability. These errors can be accounted for by using atomic correction factors as done by Dewar and Stewart (*Chem. Phys. Lett.* **111**, 416 (1984)). Like α, the second hyperpolarizability has both atomic and bond contributions and it is likely that a similar correction must be made to give accurate results. The atomic γ corrections, $\delta\gamma_i$, are treated as parameters relating the calculated and observed values of the hyperpolarizability.

$$\gamma(\text{exp}) = \gamma(\text{calc}) + \Sigma\, n_i \delta\gamma_i \ ,$$

where the sum is over types of atoms and n_i is the number of each type. As a preliminary test, data on alkanes (C_5H_{12} to $C_{10}H_{22}$) by Blaszczak and Gauden (*J. Chem. Soc., Faraday Trans. 2*, **8**, 239 (1988)) was used. The fit yielded an atomic correction for carbon of 2380 a.u. (1.20×10^{-36} esu). Using the correction gives the values in Table 1 labeled "corr. γ". These corrected values compare very well with the large basis set *ab initio* results.

It is also important in studying second hyperpolarizabilities, to explore the mechanisms for change. A major contribution to gamma is thought to come from an

extended π system (such as is found in polyacetylene). In order to examine the effect of lessening the conjugation in this systems, we have studied gamma for butadiene as a function of the CCCC torsional angle. These results are summarized in Figure 2. As can be seen, when the π-overlap is a maximum (at 0^{o} and 180^{o}) the hyperpolarizability is maximum and when the π-overlap is lest effective (at 90^{o}) the hyperpolarizability has a minimum.

Table 1. MNDO results for Polyacetylene Oligomers [$H(C_2H_2)_nH$]

All-Trans

	MNDO				*Ab Initio*	
n	γ	γ/n	γ/S	corr γ	6-31	6-31+PD
2	4255	2128		13775	1098	14846
3	30128	10043	25873	44408	9878	35118
4	103374	25844	73246	122414	40775	82212
5	249070	49814	145696	272870	114624	178443
6	483068	80511	233988	511628	253843	345721
7	810370	115767	327302	843690	476398	603537
8	1227565	153446	417195	1265645	808879	976279
9	1725875	191764	498310	1768715	1230311	
10	2295097	229510	569222	2342697	1780479	
11	2924276	265843	629179	2976636	2380428	
12	3603350	300279	679074	3660470		
13	4324536	332657	721186	4386416		
14	5079462	362819	754926	5146102		
15	5861925	390795	782463	5933325		
16	6667333	416708	805408	6743493		
17	7491513	440677	824180	7572433		

All-Cis

MNDO			
n	γ	γ/n	γ/S
2	4255	2128	
3	22137	7379	17882
4	72033	18008	49896
5	173882	34776	101849
6	346008	57668	172126
7	598899	85557	252891
8	937834	117229	338935
9	1361353	151261	423519
10	1865894	186589	504541
11	2443900	222173	578006
12	3087962	257330	644062
13	3790292	291561	702330
14	4543821	324559	753529
15	5341120	356075	797299
16	6176838	386052	835718
17	7044224	414366	867386

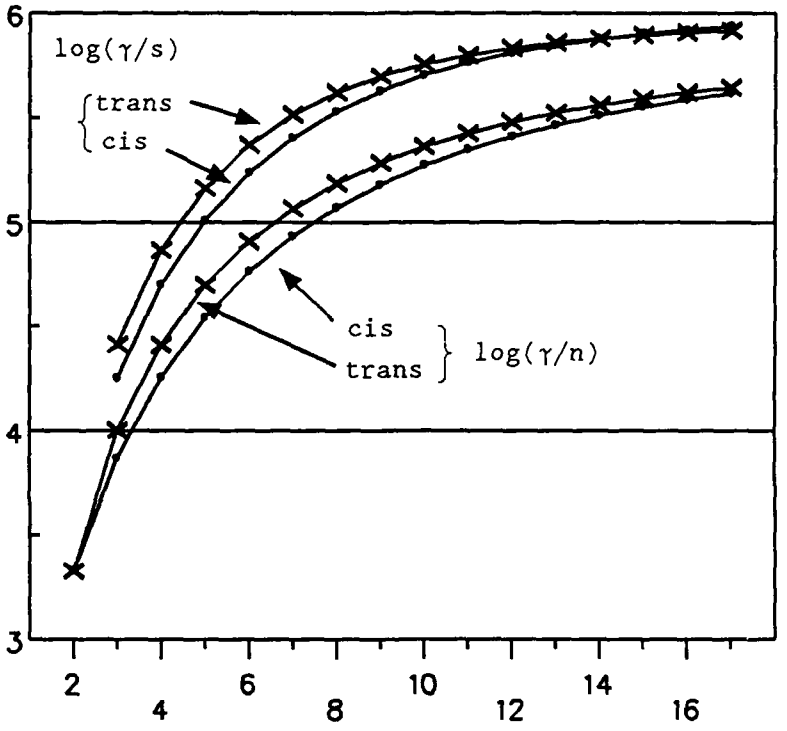

6
log(γ/s)
trans
cis
5
cis
trans
log(γ/n)
4
3
2
4
6
8
10
12
14
16
Number of subunits

Figure 1

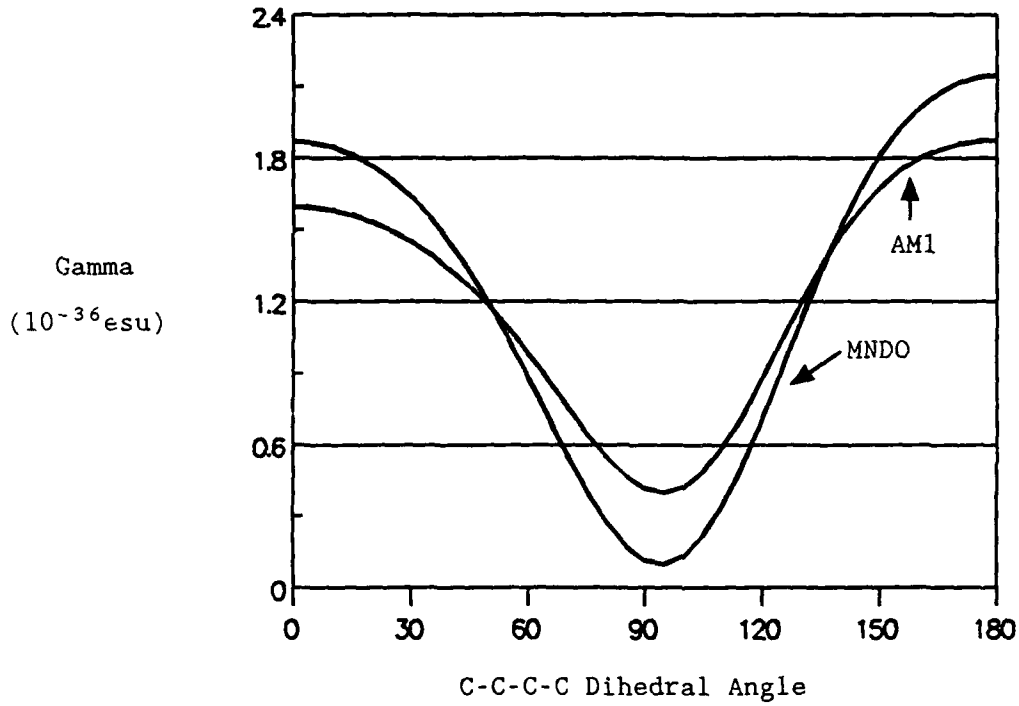

Gamma
(10^-36esu)
2.4
1.8
1.2
0.6
0
0
30
60
90
120
150
180
AM1
MNDO
C-C-C-C Dihedral Angle

Figure 2.

FINAL REPORT NUMBER 29
REPORT NOT AVAILABLE AT THIS TIME
Dr. Howard Thompson
760-7MG-071

FINAL REPORT NUMBER 30
REPORT NOT AVAILABLE AT THIS TIME
Dr. Melvin Zandler
760-7MG-092

FINAL REPORT NUMBER 31
REPORT NOT AVAILABLE AT THIS TIME
Dr. Lee A. Flippin
760-7MG-056

Final report submitted to

Universal Energy Systems

Modelling and prediction in a nonlocal turbulence model

Principal Investigator: Mayer Humi, Professor
Department of Mathematical Sciences
Worcester Polytechnic Institute
Worcester, MA 01609

USAF Contact: Don Chisholm
Chief, Meteorological Prediction Section
AFGL/LYP
Hanscom Air Force Base

Date Submitted: December 10, 1988

I. Optimal large eddy simulation

One of the fundamental problems facing turbulence simulation on a digital computer is the modeling of subgrid eddies. To overcome this difficulty large eddy simulation (LES) was proposed [1]. The essence of this parodigm [2] is to apply a filter G on the exact flow u to obtain:

$$\bar{u} = \int G(x-x')u(x')dx' \qquad (1.1)$$

(In the following we treat the one dimensional case using Burger's equation. However the extension to three dimensions is straightforward).

As a result Burger's equation becomes

$$\frac{\partial \bar{u}}{\partial t} + \overline{u \frac{\partial u}{\partial x}} = \upsilon \frac{\partial^2 \bar{u}}{\partial x^2} + \bar{F}(x,t) \qquad (1.2)$$

Substituting $u = \bar{u} + u'$ where u' are the subgrid residuals leads to

$$\frac{\partial \bar{u}}{\partial t} + \overline{\bar{u} \frac{\partial \bar{u}}{\partial x}} + \frac{\partial \bar{\tau}}{\partial x} = \upsilon \frac{\partial^2 \bar{u}}{\partial x^2} + \bar{F}(x,t) \quad . \qquad (1.3)$$

A popular model for the residuals is [3]

$$\bar{\tau} = -2\nu_T \frac{\partial \bar{u}}{\partial x} \quad , \quad \nu_T = k \frac{\partial \bar{u}}{\partial x} \qquad . \qquad (1.4)$$

Thus the two outstanding problems regarding LES are

1. The determination of the filter function G
2. The proper modeling of the subgrid residual terms $\bar{\tau}$.

The major objective of this project was to try and obtain a solution to these problems from fundamental principles rather than through numerical simulations which have limited value.

Our results regarding these problems are as follows;

1. In one dimension and for low Re (i.e. Re $\lesssim$ 200) the Gaussian filter [1] seems to be the best choice. However at higher Re ($\gtrsim$ 1000) we found that an optimal filter is given by

$$G(x) \cong 0.9\, y_0 + 0.1\, y_2$$

where

$$y_n = \alpha_n H_n(x) e^{-\frac{1}{2}x^2}$$

where $H_n(x)$ are Hermite polynomials and α_n are normalization constants.

2. To insure that the energy spectrum of $\bar{u}$ decays in the same way as u the subgrid residuals should be modelled by

$$\bar{\tau} = \bar{\bar{u}}\,\bar{u} - k\,\frac{\partial \bar{u}}{\partial x}$$

where k is a constant.

The technical details of the methodology that led to these results are given in appendix A which contains a preprint of our paper on the subject.

II. Clear air turbulence.

Current data suggests that clear air turbulence (CAT) appear in regions of strong vertical shear which is proportional to the horizontal temperature gradient. Research on this phenomena by Lilly [4] and others [5] suggests that CAT should occur when the Richardson number - Ri - is ≈ 0.25 or less. Still it is found that the correlation between theory and reality is far from being perfect especially in mountainous regions where CAT can occur even when $Ri \sim 1$. All this suggests that CAT is poorly understood on the fundamental level and a new model for this phenomena is needed. To investigate CAT more closely we wrote a finite element program using IMSL-PROTRAN to simulate the two dimensional Navier-Stokes equation in a mountainous region with various boundary and initial conditions. (This program can be extended without difficulty to the full Boussinesq system).

However each simulation of this program on the DEC-20 machine at WPI required 15-20 CPU hours for $Re \approx 100$. Simulation of higher Reynolds numbers would require even longer CPU time.
Remark: We could not use the AFGL vax cluster to this end as it does not have IMSL-PROTRAN on it.

In view of these resource requirements we had to fall back on one dimensional models (using Burger's equation with a forcing term). From these simulations we have been able to identify what we believe to be "missing" variables which might have a strong influence on the correct prediction of CAT. These are;

1. The change in the density ρ as a function of the temperature (and perhaps height). Currently all models for CAT assume ρ to be constant or hydrostatic stratification.

2. The effects of air moisture and variations in it.

3. The energy flux reaching the ground is larger at higher elevations.

It should be observed that the values of Reynolds and Richardson numbers are usually used to characterize turbulence. However, the definition of these numbers does not take into account the possible variations in ρ. This might explain in part the imperfect correlation between CAT theory and reality.

Whether a theory which take the effects mentioned above will be successful in making correct predictions regarding CAT will require a careful simulation of Boussinesq equations at least in

two dimensions. It is obvious however that such a project will require resources which go well beyond these made available to us in the present project.

References

1. See Ref. 1. of the appendix
2. See Ref. 2 of the appendix
3. See Ref. 3 of the appendix
4. D. K. Lilly et al - J. Appl. Meterol. 13 p. 488 (1974)
5. M. A. Bender et al - J. Appl. Meterol. 15 p. 1193 (1976)
6. W. Heck - Mon Weather Rev. 105 p. 1337 (1977)

   P. J. Kennedy and M. A. Shapiro - J. Atmos. Sci. 37 p. 986 (1980)

Appendix A

Integral constraints in large eddy simulation

MAYER HUMI

Mathematical Sciences Department

Worcester Polytechnic Institute

Worcester, Massachusetts 01609

ABSTRACT

We show how the optimal choice of the filter function in large eddy simulations is related to the moments of the turbulent flow.

I. Introduction

Turbulent flows contain motions in various scales. Due to this fact the computation of such flows on a finite step grid (in more than one spatial dimension) is not within the computational ability of present day computers. Large eddy simulation (LES) attempts to resolve this difficulty by proper modelling of the subgrid motions and their interaction with the resolvable flow [1].

In brief the essence of LES is to introduce a filter function [2] G(x) and define the filtered flow as

$$\bar{u} = \int G(x-x')u(x')dx \qquad (1.1).$$

By applying this filter to Navier-Stokes equations one then derives equation for $\bar{u}$ after proper modeling of the subgrid terms.

To illustrate this process we consider Burger's equation

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \upsilon\,\frac{\partial^2 u}{\partial x^2} \qquad (1.2).$$

Applying the filter operation to this equation we obtain after obvious algebra

$$\frac{\partial \bar{u}}{\partial t} + \overline{\bar{u} \frac{\partial \bar{u}}{\partial x}} + \frac{\partial \bar{\tau}}{\partial x} = \nu \frac{\partial^2 \bar{u}}{\partial x^2} \qquad (1.3)$$

where

$$\tau = u'\bar{u} + (u')^2 \qquad (1.4)$$

and u' are the subgrid residuals

$$u' = u - \bar{u} \qquad (1.5)$$

Various prescriptions are available in the literature to model $\bar{\tau}$ in terms of $\bar{u}$ e.g. Smagorinski eddy viscosity model [3] or the mixed model proposed by Bardina et al [4,5].

In one dimension these models are given respectively by [2,4]

$$\bar{\tau} = -2\nu_T \frac{\partial \bar{u}}{\partial x}, \quad \nu_T = k \frac{\partial \bar{u}}{\partial x}$$

and

$$\bar{\tau} = \bar{u}^2 - \bar{\bar{u}}^2 - 2\nu_T \frac{\partial \bar{u}}{\partial x}$$

where k is a constant

As to the filter function two principal choices are available in the literature. These are the Gaussian filter [2,5]

$$G(x) = \left[\frac{6}{\pi\Delta^2}\right]^{1/2} \exp(-6x^2/\Delta^2) \qquad (1.6)$$

and the sharp Fourier cutoff filter [5,6]

$$G(x) = \frac{2\sin(x/\Delta)}{x} \qquad (1.7).$$

It is our objective in this paper to examine the choice of the filter function in LES and show how the optimal choice of a Gaussian-like filter can be dictated by the attempt to satisfy various integral constraints on the turbulent flow and its moments.

The plane of the paper is as follows:

In section II we derive the integral constraints needed for the determination of the optimal filter function. We also show that these integral relations lead naturally to a new model for the subgrid residuals. In section III we use these results to determine the first two terms in the expansion of the optimal filter function from numerical solutions of Burger's equation using Smagorinski eddy viscosity model.

II. Orthogonal expansions and integral constraints

Motivated by the properties of the filter functions (1.6), (1.7) we place the following constraints on the optimal Gaussian-like filter function.

(1) G is a smooth function defined on $[-\infty,\infty]$.

(2) G is even.

Furthermore since (1.6) has a unit norm we require that

(3) $$\int_{-\infty}^{\infty} G(x)dx = 1$$

(4) Finally G has to be determined so that for a large class of flows the solution of the filtered equations (using some fixed subgrid model) "resembles" as far as possible the characteristics of the original flow. To accomplish this we require that the first n moments of $\bar{u}$ and u are the same.

We now discuss the implications of these requirements in detail.

To satisfy the first requirement we note that a set of orthogonal functions on $[-\infty,\infty]$ in the usual Hilbert structure of $L_2[-\infty,\infty]$ is given by [6]

$$y_n(S) = \alpha_n H_n(\zeta) e^{-\frac{1}{2}\zeta^2} \tag{2.1}$$

where $H_n(\zeta)$ are the Hermite polynomials. However since G

must have a unit norm in $L_1[-\infty,\infty]$ we must choose the normalization constants α_n so that

$$\int_{-\infty}^{\infty} y_n(\zeta)d\zeta = 1 \qquad (2.2).$$

Furthermore since $H_{2k+1}(\zeta)$, $k = 0,1,\ldots$ are odd it follows that an expansion of G in terms of $y_n(\zeta)$ must take the form

$$G(x) = \sum_{k=0}^{\infty} a_{2k}y_{2k}(\zeta), \quad \sum_{k=0}^{\infty} a_{2k} = 1, \quad a_i \geq 0 \qquad (2.3)$$

From this expression we see that from a formal point of view (1.6) represents the zero order term in the expansion (2.3) (with proper redefinition of ζ). The question that arises, therefore, naturally is whether an improvement in the optimal form of G can be obtained by additional terms in the expansion (2.3). In particular we would like to derive an algorithm to determine a first term correction to equation (1.6). To accomplish this objective we consider the moments of u and $\bar{u}$. Let u be the solution of Burger's equation on some region Ω subject to some boundary conditions. By integrating (1.2) over Ω we obtain

$$\frac{d}{dt}\int_{\Omega} u dx = \left(\nu \frac{\partial u}{\partial x} - \frac{1}{2}u^2\right)\Big|_{\partial\Omega} \qquad (2.4)$$

Similarly from (1.3) we obtain

$$\frac{d}{dt}\int_\Omega \bar{u}dx = (\upsilon \frac{\partial \bar{u}}{\partial x} - \bar{\tau})\Big|_{\partial\Omega} - \int \overline{\bar{u}\frac{\partial \bar{u}}{\partial x}}\, dx \qquad (2.5)$$

However due to the normalization of G we have

$$\int_\Omega \overline{\bar{u}\frac{\partial \bar{u}}{\partial x}}\, dx = \frac{1}{2}\int_\Omega \int_{-\infty}^{\infty} G(x-x')\frac{\partial}{\partial x'}(\bar{u}(x')^2)dx'dx = \frac{1}{2}\bar{u}^2\Big|_{\partial\Omega} \qquad (2.6)$$

Hence

$$\frac{d}{dt}\int_\Omega \bar{u}dx = (\upsilon \frac{\partial \bar{u}}{\partial x} - \bar{\tau} - \frac{1}{2}\bar{u}^2)\Big|_{\partial\Omega} \qquad (2.7)$$

Subtracting (2.7) from (2.4) we obtain that

$$\frac{d}{dt}\int_\Omega (u-\bar{u})dx = \text{Boundary terms} \qquad (2.8)$$

Similarly by multiplying equations (1.2), (1.3) by x (or more generally by x^n, $n = 2,3,\ldots$) and carrying similar algebraic computations we obtain the following equation for the first (and higher) moments of u and $\bar{u}$

$$\frac{d}{dt}\int_\Omega x(u-\bar{u})dx - \frac{1}{2}\int_\Omega (u^2-\bar{u}^2)dx + \int_\Omega \bar{\tau}dx =$$

$$= \left(x\frac{\partial u}{\partial x} - u - \frac{1}{2}xu^2\right)\Big|_{\partial\Omega} - \left(x\frac{\partial \bar{u}}{\partial x} - \bar{u} - \frac{x\tau^2}{2} - x\bar{\tau}\right)\Big|_{\partial\Omega} \qquad (2.9)$$

Moreover if we multiply (1.2), (1.3) by u, $\bar{u}$ (or u^n, $\bar{u}^n$) respectively and integrate over Ω we obtain the energy equations:

$$\frac{d}{dt}\int_\Omega u^2dx + \nu\int_\Omega \left(\frac{\partial u}{\partial x}\right)^2 = \left(\nu u\frac{\partial u}{\partial x} - \frac{u^3}{3}\right)\Big|_{\partial\Omega} \qquad (2.10)$$

$$\frac{d}{dt}\int_\Omega \bar{u}^2dx + \nu\int_\Omega \left(\frac{\partial \bar{u}}{\partial x}\right)^2 dx + \int_\Omega (\bar{\bar{u}}\bar{u}-\bar{\tau})\frac{\partial \bar{u}}{\partial x}\,dx =$$

$$= \left(\nu\bar{u}\frac{\partial \bar{u}}{\partial x} - \bar{u}\,\bar{\tau}\right)\Big|_{\partial\Omega} \qquad (2.11)$$

Our objective is to determine an optimal filter function G so that u and $\bar{u}$ have the same characteristics for a large class of flows. We can achieve this if u and $\bar{u}$ have the same moments up to some order n. However due to the "universality" requirement G must be independent of the boundary conditions of the flow. Consequently we must disregard the boundary terms in equations (2.8), (2.9) (and their analogs for $n > 1$) and choose G to minimize the absolute values of left hand side of these equations. As to the

number of moment equations to be considered a question of utility arises. On the one hand by choosing n to be large a better fit between specific u and $\bar{u}$ can be achieved. On the other hand when n is small G will have a more universal application. Due to these circumstances we shall let, in the next section $n = 1$ and determine using (2.8), (2.9) the optimal first order correction to the Gaussian filter.

As to the energy equations we consider these in the speical case where $\Omega = [0,a]$ and the boundary conditions.

$$u\big|_{\partial\Omega} = \bar{u}\big|_{\partial\Omega} = 0 \tag{2.12}$$

Under these conditions it is easy to see that an attractive model for τ which yields similar evolution equations for the energy of u and $\bar{u}$ is

$$\bar{\tau} = \bar{\bar{u}}\cdot\bar{u} - k\frac{\partial\bar{u}}{\partial x} \tag{2.13}$$

In fact with this choice of τ we can apply Poincare inequality to (2.10), (2.11) to obtain

$$\int_0^a u^2 dx \leq e^{-(\upsilon\pi^2/4a)t} \tag{2.14}$$

$$\int_0^a \bar{u}^2 dx \leq e^{-[(\upsilon+k)\pi^2/4a]t} \tag{2.15}$$

Although the prescription (2.13) for τ did not appear in the literature (to our knowledge) it is reminiscent of the mixed model of Bardina et al as $\bar{\bar{u}}$ appears naturally in the modeling of τ.

III. Optimal filter function

Our objective in this section is to determine (numerically) the first order corrections to the Gaussian filter function in the form

$$G(\zeta) = a_0 y_0(\zeta) + a_2 y_2(\zeta) \quad , \quad \zeta = \frac{2x\sqrt{3}}{\Delta} \tag{3.1}$$

where (due to normalization)

$$a_0 + a_2 = 1 \quad , \; a_i \geq 0 \tag{3.2}$$

To this end we apply the constraints given by equations (2.8), (2.9) viz. we use numerical simulations on equations (1.2), (1.3) to determine for which values of a_0 the left hand side of equations (2.8), (2.9) attains its minimum. We carry

these simulations only with Smagorinski model for the subgrid residuals.

The setting for these simulations is as follows: At first a high accuracy solution of equation (1.2) for $Re = 100, 200$ over $\Omega = [0,100]$ and time interval $[0,100]$ was obtained using a mesh with $\Delta x = 0.1$ and $\Delta t = 5.10^{-3}$. Two initial flow profiles were used,

$$u_0(x) = \sin \frac{n\pi x}{100} \quad , \quad n = 1,2 \qquad (3.3).$$

Next we simulated equation (1.3) with the same setting but with $\Delta x = 1$ for different values of a_0. Based on these simulations we conclude that at least within the context of Smagorinski subgrid model and flows with $Re = 100, 200$ the Gaussian filter function (1.6) provides superior results for $\bar{u}$ than any combination of the form (3.1) with $a_2 \neq 0$. However for $RE \approx 1000$ we obtained as optimal filter function with $a_0 = 0.9$ and $a_2 = 0.1$.

The dependence of these results on the dimension of the region as well as on the model for the subgrid residuals still remain as open questions which is under investigation at the present time.

REFERENCES

1. A review and complete list of references on the subject are given in;

   J. H. Ferziger - p. 93 in J. A. Essers, Computational methods for turbulent, transonic and viscous flow (Springer Verlag, NY 1983).

   M. D. Love and D. C. Leslie - Studies in subgrid. modeling, p. 353 in F. Durst et al (eds.), Turbulent shear flows I (Springer-Verlag, NY 1979).

2. A. Leonard - Adv. Geophys. 18 p. 237 (1974).

   B. Aupoix - in U. Frish et al - Springer notes in physics vol. 230 p. 45 (Springer Verlag, NY 1985).

3. J. Smagorinski - Mon. Weather Rev. 91 p. 99 (1963)

4. J. Bardina, J. H. Ferziger and W. C. Reynolds - AIAA Paper # 80-1357 (1980).

5. U. Piomelli, P. Moin and J. H. Ferziger - Phys. Fluids 37 p. 1884 (1988).

6. L. C. Andrew - Special functions (Macmillan, NY, 1985).

7. M. M. Denn - Stability of reaction and transport processes, Appendix C (Prentice Hall, Englewoo Cliff, NJ 1975).

FINAL REPORT NUMBER 33
REPORT NOT AVAILABLE AT THIS TIME
Dr. Steven Leon
760-7MG-036

1987-1988 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

FINAL REPORT

CO_2(001) VIBRATIONAL TEMPERATURES and LIMB-VIEW INFRARED RADIANCES UNDER TERMINATOR CONDITIONS in the 60-100 ALTITUDE RANGE

Principal Investigator:	Dr. Henry Nebel
	Associate Professor of Physics
	Alfred University
	Alfred, NY 14802
Research Location:	Air Force Geophysics Laboratory
	Optical Physics Division
Date:	22 December 1988
Contract Number:	F49620-85-C-0013/SB5851-0360

ABSTRACT

Vibrational temperature profiles as functions of altitude have been obtained for the (001) state of carbon dioxide using a non-equilibrium line-by-line infrared radiation transport code (RAD) developed at the Air Force Geophysics Laboratory. This has been done for both night-time and low-angle sunlit conditions in order to model data from the Spectral Infrared Rocket Experiment (SPIRE)[1]. Infrared radiance from the 4.3 μm bands of CO_2 in a limb view is calculated for the 60-100 km altitude range considering the line of sight to be partly sunlit and partly in darkness. Results are compared with the "night-time" SPIRE data, which were actually terminator scans.

[1]Stair, et. al., J. Geophys. Res. 90, 9763-9775 (1985).

ACKNOWLEDGEMENTS

I am grateful for the sponsorship of the Air Force Systems Command and the Air Force Office of Scientific Research, and particularly for the hospitality of the Optical Physics Division of the Air Force Geophysics Laboratory (AFGL) during the Summer of 1988 under a Research Initiation Program award. I appreciate the hospitality and guidance of Dr. Ramesh D. Sharma of AFGL with whom I worked closely during my time at AFGL. I thank Drs. Richard Picard and Jeremy Winick of AFGL for a number of useful discussions and for providing information on the SPIRE data set, and Dr. Robert Joseph of Arcon Corporation for assistance in revising the computer code. I thank Universal Energy Systems, Inc. for administration of the Research Initiation Program.

Finally, I thank Dr. Peter Wintersteiner of Arcon Corporation for many useful discussions and suggestions, and for assistance in implementing the infrared radiance code. This research would not have been possible without his assistance.

INTRODUCTION

There has been increasing interest recently in radiative transfer in the infrared spectral region under non-equilibrium conditions in the atmosphere (1-5). Under these conditions, local thermodynamic equilibrium (LTE) may not be assumed to apply, i.e. collisions among molecules are not frequent enough to bring a parcel of air into equilibrium before radiative deexcitation occurs. This is generally the situation in the earth's upper atmosphere (above 60 km). An infrared radiance computer code has been developed at the Air Force Geophysics Laboratory (AFGL) by Dr. Ramesh Sharma of AFGL and Dr. Peter Wintersteiner of Arcon Corporation. This code treats absorption, emission, and transmission of infrared radiation through the atmosphere under non-equilibrium conditions. One component of the code (RAD) calculates excited state population densities and the corresponding vibrational temperature profiles assuming various mechanisms of excitation and deexcitation, and also assuming the population densities are constant in time. Another component of the code (NLTE) uses the vibrational temperature profiles to calculate total integrated band radiance for various viewing geometries. All calculations are done on a line-by-line basis, whereas most previous calculations of this type were based on band models (1,2,5). I spent the summers of 1986 and 1987 at AFGL under the sponsorship of the US Air Force Summer Faculty Research Program. This time was spent applying the infrared radiance code to the asymmetric stretch mode of excitation of carbon dioxide (4.3 micron band) under quiescent night-time conditions (6,7). Limb-view integrated radiances were calculated in order to compare with night-time data from the Spectral Infrared Rocket Experiment, or SPIRE (9). I received a Research Initiation Program (RIP) award for 1988 which allowed me to return to AFGL for the summer of 1988 during which time I applied the code to the same band of carbon dioxide under terminator conditions, i.e.

part sunlit, part darkness. This was done in an attempt to model the actual experimental conditions of the night-time SPIRE data, as will be explained later in this report.

THEORETICAL DEVELOPMENT

The level scheme of interest for the 4.3 μm band of CO_2 is shown in Fig. 1. The numbers in parentheses at each CO_2 level refer to numbers of vibrational quanta in the three possible modes of oscillation (ν_1, ν_2, ν_3) of the molecule. The (001) state (one quantum of ν_3 vibration) is responsible for the 4.3 μm band. There is rapid collisional transfer between this state and the first excited vibrational state of N_2; thus the populations of both these states are treated as unknowns in the calculation. The mechanisms assumed for excitation and deexcitation of these two states are shown in Fig. 2. Rate constants for these reactions have been taken from reference 5. Each of the two excited states is assumed to have a population density which is constant in time. Thus the sum of the excitation rates is set equal to the sum of the deexcitation rates for each state. The result is two coupled linear equations in the population densities of the excited states in question. These equations are combined to form a 2x2 matrix equation which is solved by the computer code yielding the population densities of the $CO_2(001)$ state and the $N_2(1)$ state. The vibrational temperatures are then calculated by means of the formulas

$$[CO_2(001)]/[CO_2(000)] = \exp(-h\nu/kT_{001})$$

and

$$[N_2(1)]/[N_2(0)] = \exp(-h\nu/kT_{N2})$$

where [ ] represents a population density, T_{001} and T_{N2} are the vibrational temperatures of the $CO_2(\upsilon 01)$ state and the $N_2(1)$ state respectively, h and k are the Planck and Boltzmann constants respectively, and ν is the frequency associated with the transition in each case. These equations define the vibrational temperatures. The calculation may be done for each altitude required for a particular problem, resulting in a vibrational temperature profile as a function of altitude. This vibrational temperature profile may then be used to calculate total band radiance in a limb view, the geometry of which is shown in Fig. 4. The program calculates integrated radiance in a limb view for each spectral line within the band for selected tangent heights (see Fig. 4), and then sums the results to obtain total band radiance. The results may then be compared with experimental data obtained from a rocket or a satellite.

RESULTS

A. Vibrational Temperatures

Night-time vibrational temperature profiles as functions of altitude for the asymmetric stretch (ν_3) mode of excitation of carbon dioxide (4.3 μm band) have been obtained using a component of the infrared radiance code called RAD. This program calculates population densities of excited states and the correspor ling vibrational temperatures as outlined above. The subroutine which actually calculates the vibrational temperatures was modified to solve the 2x2 matrix equation discussed above, for previously the code had only been applied to the 15 μm band of CO_2. Program RAD has been updated during 1988 so that it performs the calculations of vibrational temperatures much faster than was previously possible. Results have been obtained under night-time conditions for the principal isotopic form of carbon dioxide ($^{12}C^{16}O_2$)

labeled 626, as well as for the three most important minor isotopic forms ($^{13}C^{16}O_2$, $^{16}O^{12}C^{18}O$, $^{16}O^{12}C^{17}O$) labeled 636, 628, and 627 respectively. These results are shown in Fig. 3 for the 50-150 km altitude range.

B. Limb View Integrated Radiances

Total infrared radiance from the 4.3 μm band in a limb view (see Fig. 4) has been calculated using a component of the infrared radiance code called NLTE (8). This program uses the vibrational temperature frofiles as input, and calculates integrated radiance in a limb view for each spectral line within the band for selected tangent heights (see Fig. 4). It then sums the results to obtain total band radiance for each selected tangent height. This has been done for the fundamental band using the four isotopic forms referred to above, and also for the most important "hot bands" (higher order transitions) using the two strongest isotopic forms. The sum of these radiance results is shown in Fig. 5. Also shown in this figure are experimental integrated radiance values obtained from the Spectral Infrared Rocket Experiment, or SPIRE (9). As can be seen in Fig. 5, the calculated radiance values are smaller than the measured values above 60 km. The reason for this is that the limb view lines-of-sight used in the SPIRE mission to obtain "night-time" radiance were not totally in darkness. The corresponding instrument scans were actually "terminator" scans (part darkness, part sunlit). The rocket itself was in sunlight at the time of the measurements; therefore a portion of each line-of-sight was also in sunlight, as can be seen in Fig. 6. Absorption of sunlight will increase the number of CO_2 molecules in excited states, thereby increasing the measured radiance in a limb view. Thus it is expected that the measured radiance values would be larger than values calculated without assuming absorption of sunlight.

C. Terminator Conditions

To account for the "terminator" scans, total integrated radiance in a limb view has been calculated assuming the line-of-sight is partly in darkness and partly sunlit with a low angle sun. Solar absorption was added as an additional excitation mechanism in the calculation of vibrational temperature profiles for the $CO_2(001)$ state. Results are shown in Fig. 7 for the four isotopic forms referred to above. To calculate total integrated radiance in a limb view for the terminator scans, the geometry of the SPIRE experiment was analyzed to determine what portions of the lines-of sight were sunlit and what portions were in darkness. A modified form of program NLTE was used which calculates integrated radiance in a limb view assuming night-time vibrational temperatures for part of the line-of-sight and low angle sunlit vibrational temperatures for the remainder of the path. The results of these calculations for the fundamental band including all four isotopic forms of CO_2 referred to above are shown in Fig. 8, again along with the corresponding data from the SPIRE mission. It can be seen in Fig. 8 that the agreement with experiment is much better than in Fig. 5.

RECOMMENDATIONS

A. Terminator Conditions

Several refinements are still necessary for this problem. The calculation of vibrational temperature profiles assumes laterally homogeneous layers in the atmosphere. This will not be the case under low angle sunlit conditions because for any particular layer, the portion of the layer on the sun side of the tangent point (see Fig. 4) receives sun rays at a different angle from the portion on the opposite side.

Also, the sun was assumed to be above the horizon (solar zenith angle = 88°) in the calculation of sunlit vibrational temperature profiles, when in reality it was somewhat below the horizon as may be seen in Fig. 6. It is recommended that methods be sought to improve upon these approximations.

In the calculations described here, it has been assumed that the hydroxyl radical (OH) plays no role in the kinetic equations for excitation and deexcitation of the $CO_2(001)$ state. It is well known that hydroxyl exists in the upper atmosphere (10), and it has been suggested (11) that interactions involving the OH radical will have a significant effect on $CO_2(001)$ vibrational temperatures. It is recommended that the effect of hydroxyl be investigated by including it in the kinetic equations and re-calculating vibrational temperature profiles and integrated radiances. Comparison with experimental results can then be made as before.

B. Day-time Conditions

It is recommended that these calculations be extended to treat the case of full day-time conditions. This will involve modifying the kinetic equations to include solar pumping at higher sun angles, determining vibrational temperature profiles which correspond to these sun angles, and using the resulting profiles to calculate total band radiance in a limb view for various tangent heights. The results can be compared with data from the SPIRE mission which took several fully sunlit scans. Data from these scans were converted to total day-time band radiance in a limb view as a function of tangent height (Reference 9, Fig. 21).

SUMMARY AND CONCLUSIONS

The US Air Force Research Initiation Program (sponsored by AFOSR, conducted by UES) has been very successful in allowing me to continue research begun while on the Summer Faculty Research Program during the summer of 1987. The work described in this report, namely calculation of $CO_2(001)$ vibrational temperatures and limb-view integrated radiances under terminator conditions in the upper atmosphere for comparison with data from the SPIRE mission, carries out exactly what I recommended in my Final Report for the SFRP in 1987. Most of the work was performed during the summer of 1988 while in residence at the Air Force Geophysics Laboratory, Hanscom AFB, Mass. In addition, the RIP award provided funds for two hours of released time from teaching during each of the semesters Spring 1988 and Fall 1988. During the spring the time was spent transferring computer programs from the Control Data CYBER computer at AFGL to the Digital VAX machine at Alfred University. This allows me to run the programs at Alfred as well as at AFGL. During the fall semester the time was spent preparing a paper for presentation at the Fall Meeting of the American Geophysical Union, and also writing this final report. The presentation at the AGU Meeting (Nebel, et. al., Eos Transactions, AGU, 69, 1346, (1988)) summarized the research described in this report. The continuity fostered by these research programs has been very beneficial to me and to my research colleagues at AFGL. I hope to continue this research under future sponsorship of AFOSR. Given such an opportunity, my first priority would be to carry out the recommendations listed earlier in this report.

REFERENCES

1. Kumer, J. B., and T. C. James, "CO2(001) and N2 Vibrational Temperatures in the $50 \lesssim z \lesssim 130$ km Altitude Range", J. Geophys. Res. 79, 638-648 (1974).

2. Shved, G. M., G. I. Stepanova, and A. A. Kutepov, "Transfer of 4.3 μm CO2 Radiation on Departure from Local Thermodynamic Equilibrium in the Atmosphere of the Earth", Izvestiya, Atmospheric and Oceanic Physics, 14, 589-596 (1978).

3. Sharma, R. D., and P. P. Wintersteiner, "CO2 Component of Daytime Earth Limb Emission at 2.7 Micrometers", J. Geophys. Res. 90, 9789-9803 (1985).

4. Solomon, S., J. T. Kiehl, B. J. Kerridge, E. E. Remsberg, and J. M. Russell III, "Evidence for Nonlocal Thermodynamic Equilibrium in the ν_2 Mode of Mesospheric Ozone", J. Geophys. Res. 91, 9865-9876 (1986).

5. López-Puertas, M., R. Rodrigo, J. J. López-Moreno, and F. W. Taylor, "A non-LTE radiative transfer model for infrared bands in the middle atmosphere. II. CO2(2.7 and 4.3 μm) and water vapor (6.3 μm) bands and N2(1) and O2(1) vibrational levels", J. Atmos. Terr. Phys. 48, 749-764 (1986).

6. Nebel, H., "CO2(001) Vibrational Temperatures in the 50 to 150 km Altitude Range", Final Report/ 1986 USAF-UES Summer Faculty Research Program (1986).

7. Nebel, H., "Night-Time CO2(001) Vibrational Temperatures and Limb-View Integrated Radiances in the 50 to 150 km Altitude Range", Final Report/ 1987 USAF-UES Summer Faculty Research Program (1987).

8. Wintersteiner, P. P., and R. D. Sharma, "Update of an Efficient Computer Code (NLTE) to Calculate Emission and Transmission of Radiation Through Non-Equilibrium Atmospheres", AFGL-TR-85-0240 (1985).

9. Stair, A. T., R. D. Sharma, R. M. Nadile, D. J. Baker, and W. F. Grieder, "Observations of Limb Radiance With Cryogenic Spectral Infrared Rocket Experiment", J. Geophys. Res. 90, 9763-9775 (1985).

10. Baker, D. J., T. Conley, and A. T. Stair, "On the Altitude of the OH Airglow", Eos Trans. AGU, 58, 460 (1977).

11. Kumer, J. B., A. T. Stair, N. Wheeler, K. D. Baker, and D. J. Baker, "Evidence for an $OH^* \xrightarrow{VV} N2^* \xrightarrow{VV} CO2(\nu 3) \longrightarrow CO2 + h\nu(4.3\,\mu m)$ Mechanism for 4.3 μm Airglow", J. Geophys. Res. 83, 4743-4747 (1978).

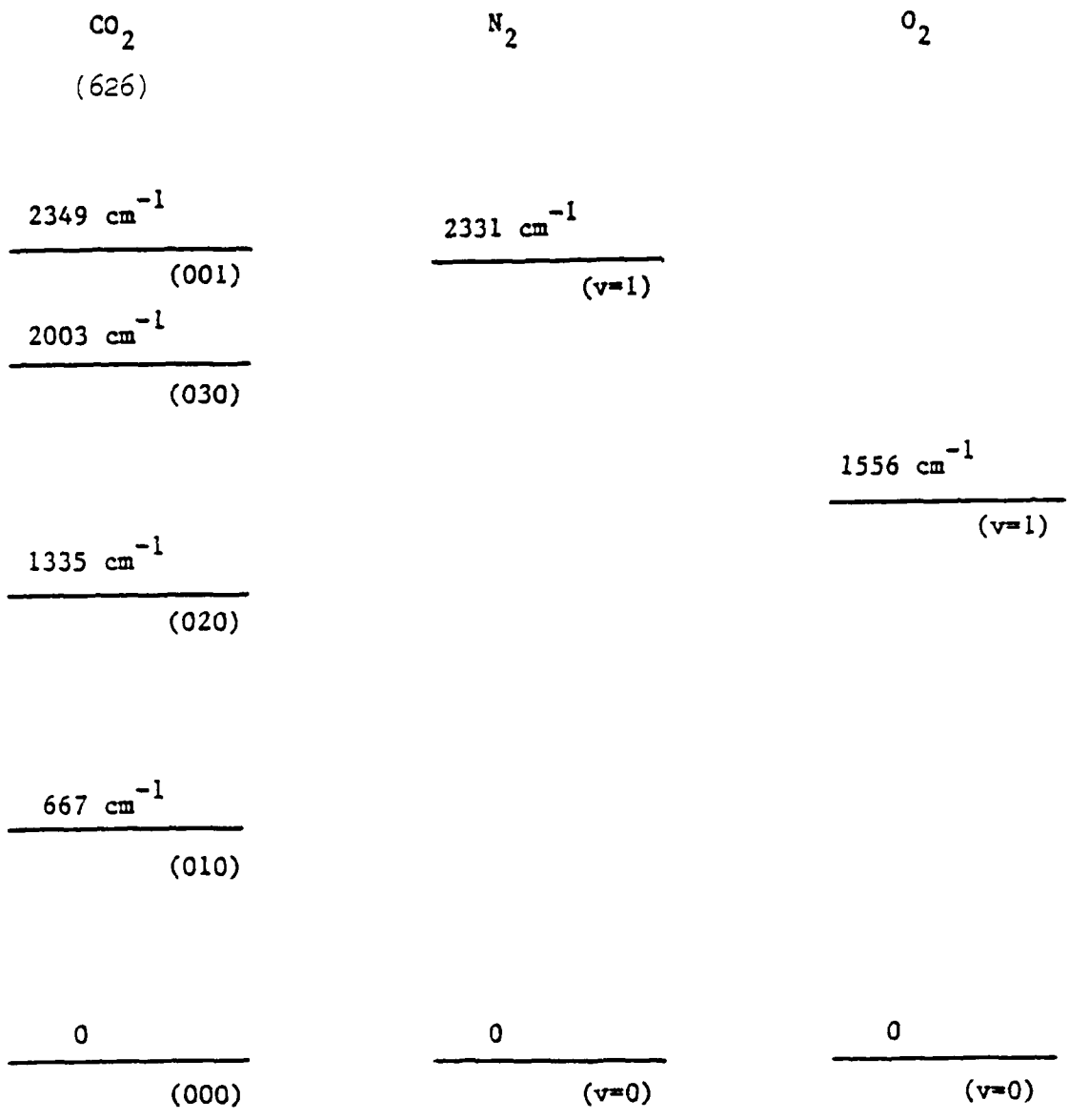


Fig. 1. Low-lying vibrational levels of CO_2, N_2, and O_2.

For the CO_2 state we have the following processes:

Spontaneous emission:

$$CO_2(001) \longrightarrow CO_2(000) + h\nu$$

Earthshine pumping:

$$CO_2(000) + h\nu \longrightarrow CO_2(001)$$

V-V transfer with N_2:

$$CO_2(001) + N_2(0) \rightleftharpoons CO_2(000) + N_2(1) + 18cm^{-1}$$

V-T transfer to $CO_2(030)$:

$$CO_2(001) + M \rightleftharpoons CO_2(030) + M + 346cm^{-1}$$

V-T transfer with oxygen atoms:

$$CO_2(001) + O \rightleftharpoons CO_2(000) + O + 2349cm^{-1}$$

V-V transfer with O_2:

$$CO_2(001) + O_2(0) \rightleftharpoons CO_2(010) + O_2(1) + 128cm^{-1}$$

For the N_2 state we have the following processes:

V-V transfer with CO_2:

$$CO_2(001) + N_2(0) \rightleftharpoons CO_2(000) + N_2(1) + 18cm^{-1}$$

V-V transfer with O_2:

$$N_2(1) + O_2(0) \rightleftharpoons N_2(0) + O_2(1) + 775cm^{-1}$$

V-T transfer with M (N_2 or O_2):

$$N_2(1) + M \rightleftharpoons N_2(0) + M + 2331cm^{-1}$$

V-T transfer with oxygen atoms:

$$N_2(1) + O \rightleftharpoons N_2(0) + O + 2331cm^{-1}$$

Fig. 2. Night-time mechanisms for $CO_2(001)$ and $N_2(1)$ states.

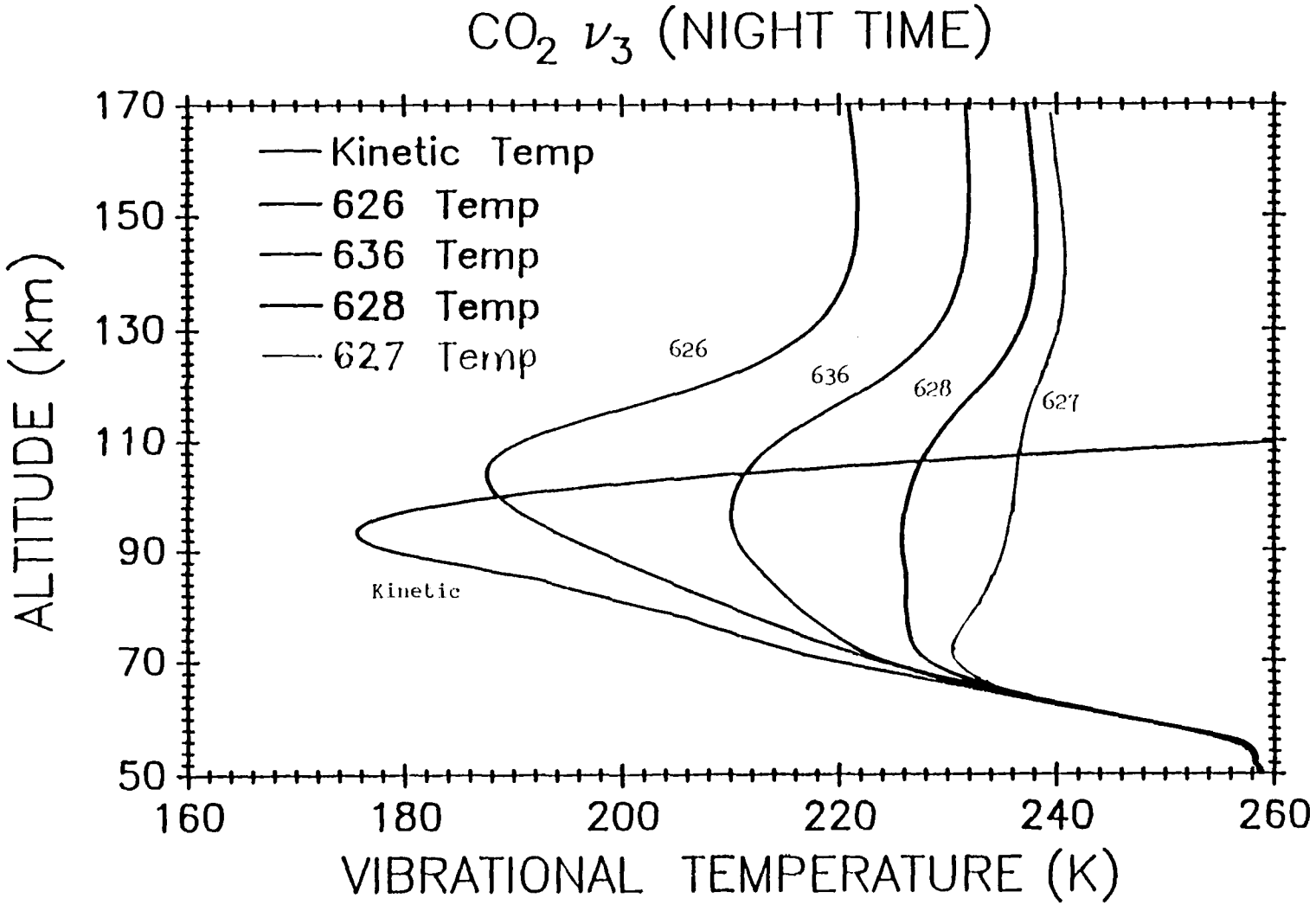


Fig. 3. Vibrational temperatures for night-time conditions.

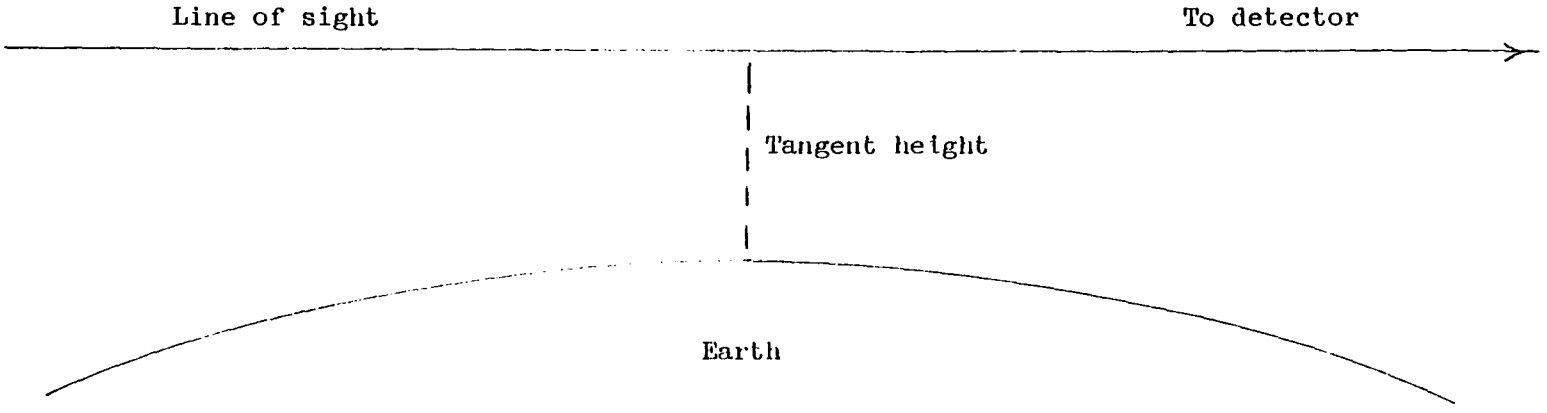


Fig. 4. Limb viewing geometry.

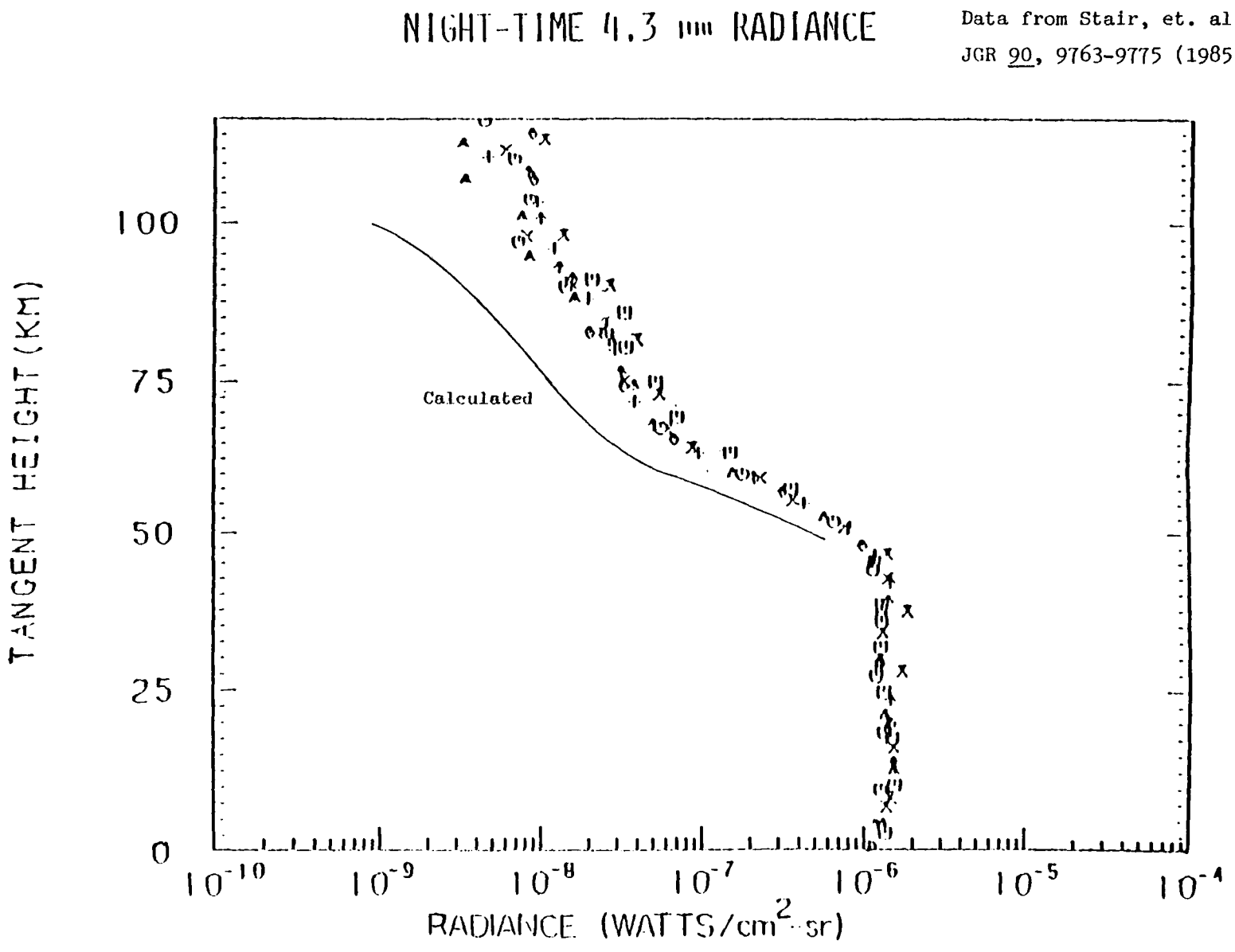


Fig. 5. Integrated radiance in a limb view. Calculated: Night-time conditions.

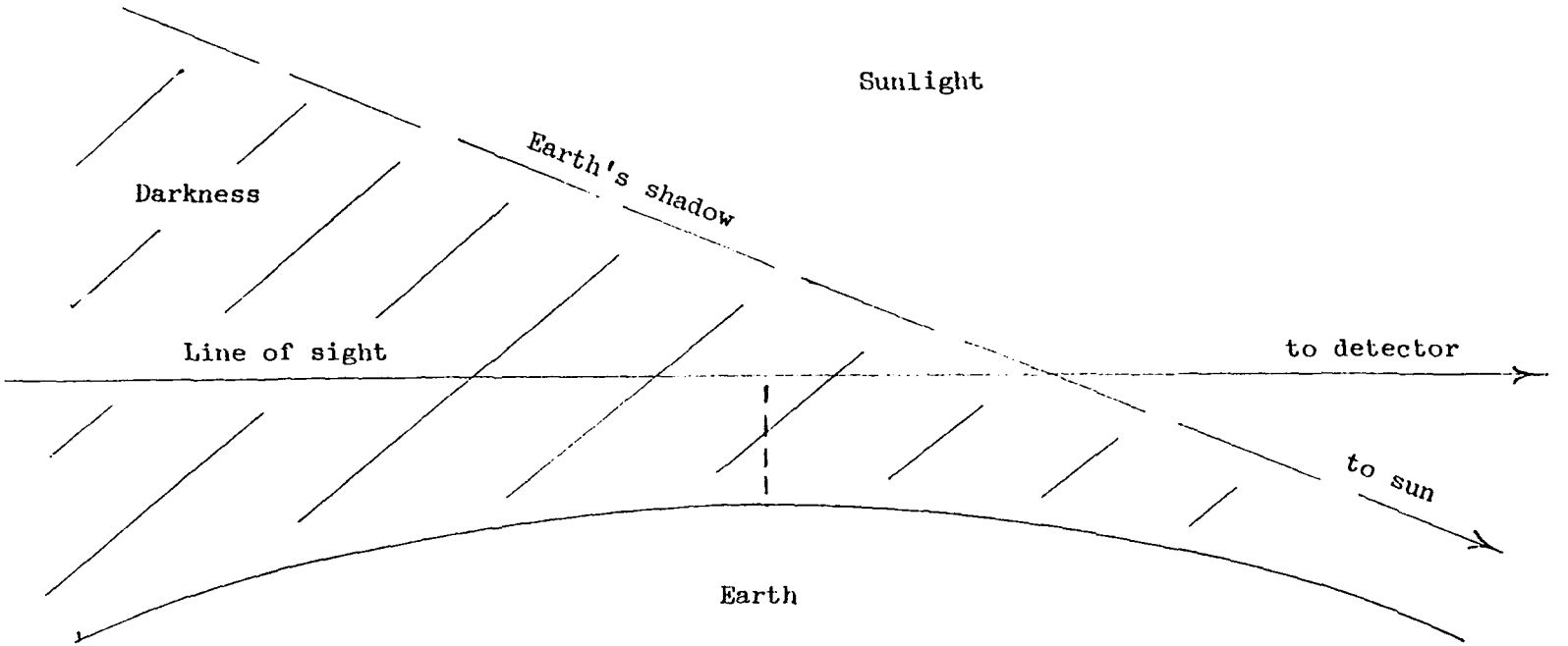


Fig. 6. Limb view near terminator.

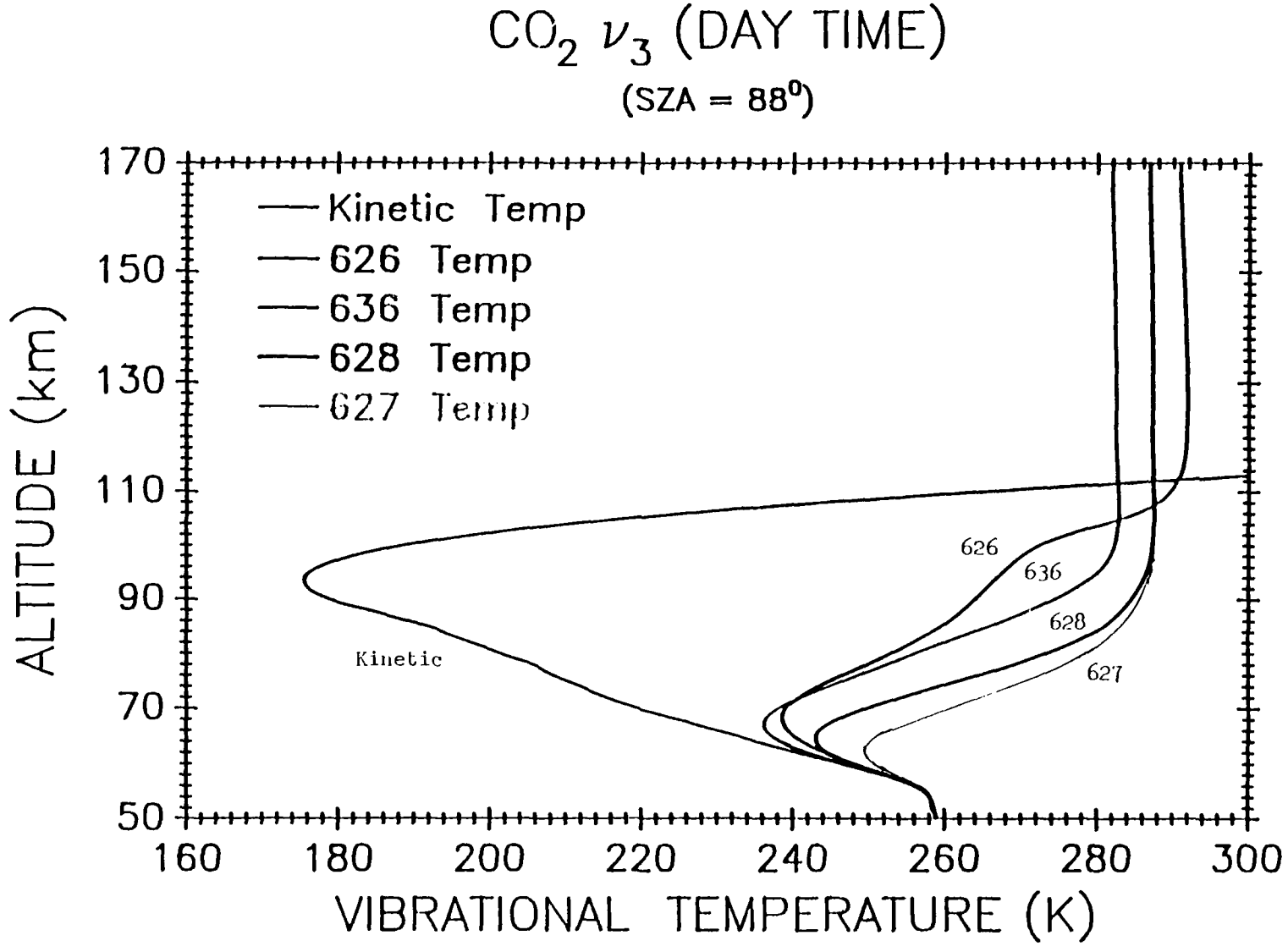


Fig. 7. Vibrational temperatures for low-angle sunlit conditions.

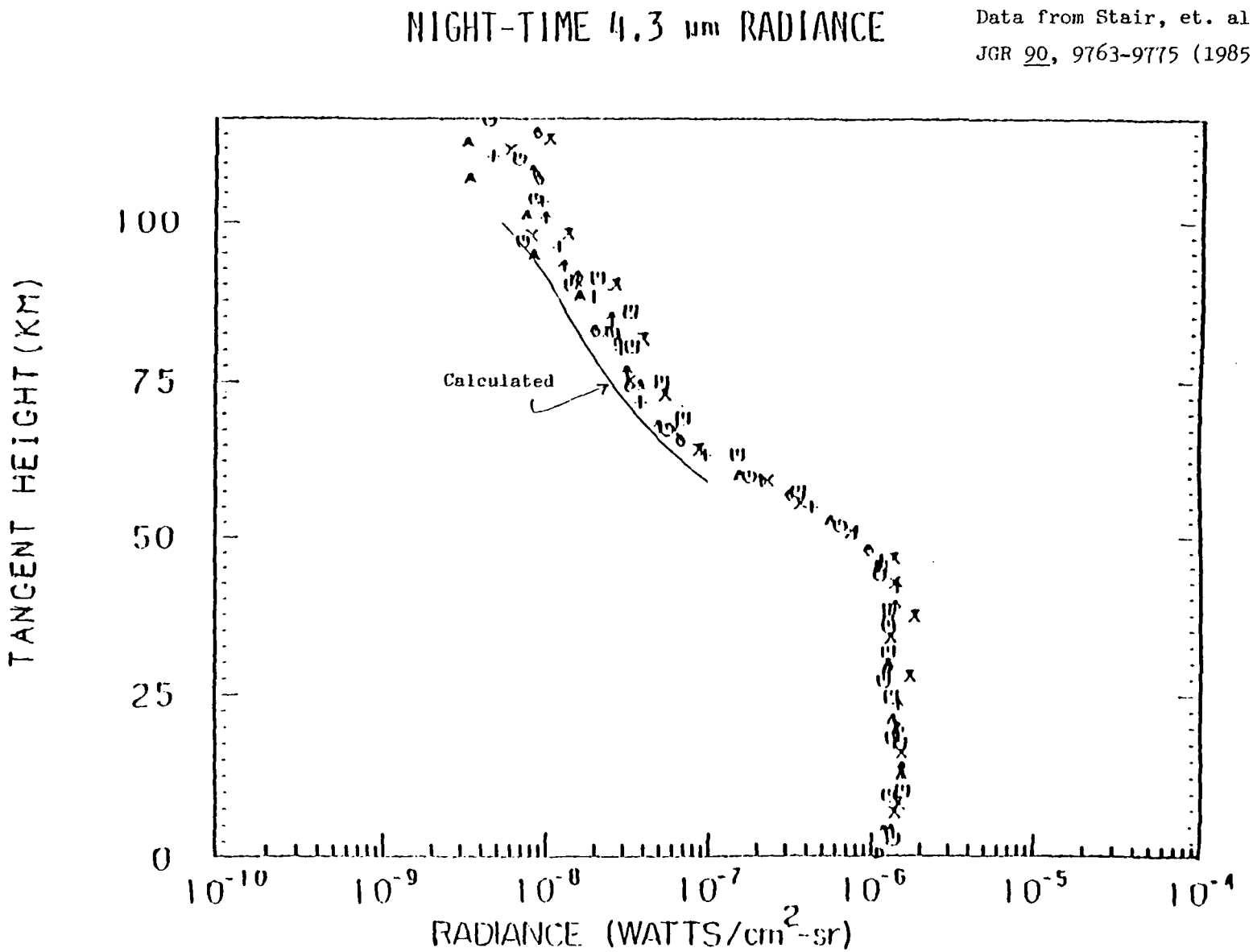


Fig. 8. Integrated radiance in a limb view. Calculated: Terminator conditions.

COMPARISON OF SSM/I RAINRATES AND SURFACE WINDS WITH THE CORRESPONDING CONVENTIONAL DATA IN THE NORTH WEST PACIFIC TYPHOONS

Gandikota V. Rao

Department of Earth and Atmospheric Sciences

Saint Louis University

St. Louis, Missouri 63103

Research performed under Research Initiation Award under contract Number F49620-85C-0013 with the Air Force Office of Scientific Research, Bolling AFB, D.C., through the Universal Energy Systems

Abstract

Rainrate data from the recently (1987) launched DMSP SSM/I (Special Sensor Microwave Imager) was acquired for three maptimes (on 25 and 26 October 1987) for Typhoon Lynn when that storm was close (about 300 km) to the southern Taiwan. This particular cyclone was chosen because it was nearly stationary and slowly deintensifying at that time and was also under radar surveillance. The areas close to (50 km) and far from the radar (more than 250 km) were excluded from the radar analysis. The 11 cm weather radar at Kaohsiung was able to resolve fine precipitation features (4 km scale). On the other hand the SSM/I 37 GHz footprint has a spatial resolution of about 37 x 28 (sq. km) - good for space borne platform but could resolve only coarse features. This comparison of the radar and SSM/I rainrates revealed that the SSM/I resolved only the gross features of the typhoon precipitation. Although individual rainbands were not resolved properly by the SSM/I heavy precipitation areas coincided with those so inferred by radar.

The SSM/I winds and the conventional (ship, coastal, island and reconnaissance) winds were compared for typhoon Thelma on 13 and 14 July 1987 and for typhoon Lynn for 25 and 26 October 1987 in the North West Pacific. The correlation was high (0.7). The SSM/I wind speeds agreed well with the surface truth in the range 15-22 m s^{-1}. The SSM/I winds in general overestimated in the lower range of wind speeds and underestimated in the upper range.

1. Introduction

According to Liou (1980) microwave radiation would reveal the surface characteristics faithfully because the radiation emanating from the surface does not suffer much attenuation from absorption from any gas in the atmosphere. For example only a weakly absorbing pressure broadened 22.235 GHz water vapor line and a cluster of oxygen lines centered around 60 GHz exist in the microwave spectrum. Low transmissivities also exist at 118.75 GHz and 183 GHz (see Liou's Figure 7.11) but these are not very broad. Consequently relatively unattenuated radiation is caught by the satellite.

Rao (1984) in a comprehensive review article pointed out the advantage of microwave radiometry in the measurement of rainfall over the oceans. The satellites NIMBUS 5 and 6 used the Electrical Scanning Microwave Radiometers (ESMR) and measured the 19 GHz (linearly polarized) and 37 GHz (dual polarized) respectively to infer the oceanic rainfall. This review article has an extensive list of references. Therefore these references are not listed here.

Bunting and Hardy (1984) discussed how the brightness temperature of the earth's surface depends on surface type and condition. For example the ocean emissivity depends on wind speed and land emissivity on the soil temperature. They further commented how the Special Sensor Microwave (SSM/I) Imager aboard a DMSP satellite would remotely sense the surface wind speed and rainrates over the vast expanse of the oceans.

As anticipated in June 1987 a DMSP satellite was launched with the SSM/I aboard. This particular microwave imager has four channels with

vertical and horizontal polarization detection capability. The chief objective of this imager was to provide surface wind and precipitation patterns over the oceans.

The primary objective of this study is to compare the microwave derived precipitation fields and surface wind fields pertaining to some typhoons with radar derived precipitation and conventional wind fields of the corresponding typhoons. Such a comparison is intended to be qualitative only. It may be mentioned that Lovejoy and Austin (1980) and Raschke and Ruprecht (1981) compared radar rainrates obtained in the GATE (Garp Atlantic Tropical Experiment) with those derived by the satellite microwave radiometers. Lovejoy and Austin (1981) attached only + or - 70% accuracy to the microwave measurements because of the unknown amount of cloud water and the depth of the rain layer. Although the lack of knowledge of these valuable parameters continues to be a problem it is hoped that the demonstration, in our case, that the SSM/I data are reasonably accurate encourages the employment of the data for the prediction of typhoon intensity and track speed. Our empirical study is different from that of Spencer et al. (1988) who compared the polarized corrected temperatures derived from the 85 GHz against the radar imagery of an eastern Atlantic rain storm (summer).

2. Some of the DMSP SSM/I Chief Characteristics

The DMSP (Block 5D-2) spacecraft was launched in a circular sun synchronous near polar orbit at an altitude of 833 km with a period of 102 minutes. The SSM/I swath width is 1400 km and results in a dense coverage on successive days. The SSM/I is a seven channel, four fre-

quency linearly polarized passive microwave radiometric system. Table 1 shows the details of the channel frequency and the Effective Field of View (EFOV).

Holinger, et al. (1987) discussed in detail the footprint geometry of the 37 GHz and 85.5 GHz frequencies. It appears from Table 1 that the elliptical areas of 18.5 km major axis and 14 km minor axis contain the 37 GHz information such as rainrates or winds. The 85.5 GHz contains information on a finer scale (15 km) but this information is available only in terms of brightness temperatures but not (customarily) in terms of precipitation rate or wind speed. Thus it is fair to say that the SSM/I rainrates and wind speed are available at distances of 35 km. This scale of resolution is important when a comparison of the radar rainrates and the satellite rainrates is made.

3. Radar Rainrates

In comparing the SSM/I rainrates against the radar rainrates some discussion of the derivation of radar rainrates is necessary. Most investigators of radar rainrates consider an empirically derived relationship of the form

$$Z = AR^b \qquad (1)$$

Table 1. Showing some characteristics of the SSM/I suite

Channel freq.	Polarization	EFOV on earth surface km along track	cross
19.35	V	69	43

19.35	H	69	43
22.235	V	50	40
37.0	V	37	28
37.0	H	37	28
85.5	V	15	13
85.5	H	15	13

relating reflectivity factor Z to the rainrate R. Battan (1973) has tabulated over sixty Z-R relationships based on world wide measurements.

For reliable precipitation estimates the majority of the radar volume should lie entirely below the freezing level. This means with a 2° elevation angle the maximum range for radar estimation of rain is 175-225 km assuming that the freezing level is located at 5 km.

Standard books on radar applications (e.g., Rogers, 1976) discuss how Z depends on the dropsize distribution and how it is sensitive to the large drop component of such a distribution. For a Marshall-Palmer (MP) distribution of raindrops extending from Zero diameter to infinity the reflectivity factor is given by

$$Z = \frac{No\ (6!)}{(41)^7} R^{1.47} \qquad (2)$$

The empirical data on Z and R for rain shows the relationship

$$Z = 200R^{1.6} \qquad (3)$$

holding reasonably well.

Recently Jorgensen and Willis (1982) established a Z-R relationship

$$Z = 300R^{1.35} \quad (4)$$

based on aircraft data from four flights into three storms at three altitudes. The composite Z-R relationship is shown to be good both for convective and stratiform rain regions. Jorgensen and Willis (1982) reviewed various Z-R relationships and found the MP relationship (Eq. 3) to yield higher rainrates in the lower Z's than what Eq. (4), called JW, would give. Most of the reflectivities in hurricanes are traditionally lower than 48 dBz. Thus near the lower end e.g., 25 dBz the MP relationship would infer better rainrates than the JW relationship.

In the following study the spatial distribution of rainrates is displayed (in Figs. 3, 5 and 7) using the JW formulation. While developing correlation coefficients and regression estimates both formulations are used.

4. Comparison of the SSM/I and Radar Rainrates

The precipitation associated with mature typhoons as inferred by the SSM/I is compared with the radar derived precipitation. The radar is stationed in Kaohsiung, Republic of China. Table 2 shows the specifications of this radar. Since the SSM/I started to function late in June 1987 tropical cyclones that occurred in July 1987 and thereafter are of interest in this study. Ideally the tropical cyclones should travel slowly and not experience any rapid changes in intensity. They should be about 300 km away from the radar toward the ocean. The bulk of the tropical cyclone circulation should be near the satellite subtrack point. Only a few storms are available in 1987 satisfying these conditions. For July and October 1987 two typhoons: Thelma and Lynn

were studied. Of these two the typhoon Lynn had more data coverage and therefore was discussed first.

Table 2. Specification of the Kaohsiung radar

Specification	Value
Wavelength	11.1cm
Frequency	2900 MHz
Peak transmitting power	500 kw
Pulse rep. freq.	164 pps
Beam width	2.25 degrees

Fig. 1 shows the track of typhoon Lynn during October 15-27, 1987. Of current interest are the two days 25 and 26 when Lynn was close to Kaohsiung.

Fig. 2 shows the SSM/I rainrates for 1035 UTC 25 October. Most of the rain is concentrated in bands east and south east of the storm center. The 8 mm h^{-1} isopleth covers an elliptical area of 50 km (major axis) and 15 km (minor axis). Gradients of precipitation are packed south east of Lynn.

Fig. 3 shows the radar precipitation for 1100 UTC 25 October. A circular area of radius 100 km with Kaohsiung as center is blocked. This is interpreted as sea clutter. Three prominent rainbands are noticed in this figure. It is noteworthy that the SSM/I rainrates (e.g., 8 mm h^{-1} isopleth) are north-south oriented while the radar rainrates have a southeast northwest orientation. Unlike the twin maxima in Fig. 3 only one maximum near 21N and 120E was shown by the SSM/I. This lack of agreement between the two figures is likely due to the

differences in time and scales of measurement (16 square km for radar versus 850 square km for SSM/I).

Figs. 4 and 5 show the corresponding maps for 22 UTC 25 October 1987. The lack of proper curvature to the rainbands and depiction of one rainband near 21.5 N and 119.5 E in Fig. 4 in lieu of two in Fig. 5 (respectively at 21.2N, 119.4E and at 21N and 120E) are the discrepancies most noticed. The isolated band to the southwest of Taiwan at 22.2N and 119E appears alike in both Figs. 4 and 5.

Figs. 6 and 7 show the SSM/I rainrates and the corresponding radar rainrates for 1000 UTC 26 October. The SSM/I pattern developed the proper curvature but underestimated the rain amounts in general.

Fig. 8 shows the linear regression line between the SSM/I and radar rainrates for the combined three maptimes. The radar rainrates were developed using the JW formulation $Z = 300R^{1.35}$. From the regression it is found that when the radar is showing a rainrate of 2 mm h^{-1} the SSM/I is producing 3.8 mm h^{-1} - an obvious overestimate. As the radar rate goes up this overestimate decreases. The correlation (0.49) coefficient is low. Such a relatively poor correlation resulted because at low radar rainrates the SSM/I was registering rainfall over a wide range of rates. In the middle range i.e., 3 to 5 mm h^{-1} (radar) the scatter was confined to 5 to 7 mm h^{-1} (SSM/I) indicating a better agreement. With the MP (Fig. 9) formulation the correlation improved slightly. As was found earlier by Jorgensen and Willis (1982) the MP formulation yields slightly higher estimates of radar precipitation for lower reflectivity values than the JW formulation. Therefore in the lower range of precipitation values the MP fared better in terms of correlation coefficient.

5. Comparison of the SSM/I and Conventional Winds

Fig. 10 shows the surface winds as inferred by the SSM/I for 22 UTC 24 October 1987. An envelope of 30 m s^{-1} surrounds the typhoon center. In general strong winds (15 m s^{-1}) are observed over the map. Fig. 11 shows the surface winds as observed by the ships (+ or - 3h of SSM/I time) and some islands. Hsu (1986) suggested a formula connecting the land and sea observations:

$$U_{sea} = 1.62 + 1.17\ U_{land}.$$

This formula permits a comparison of the SSM/I observation in the vicinity (50 km or more) of an island with the neighboring ship observations.

A comparison of Figs. 10 and 11 shows that a fair agreement exists between the SSM/I and the observed winds. The winds northeast of Taiwan appear to be exaggerated by the SSM/I.

Table 3 shows the maptimes for which wind data were gathered for developing a correlation coefficient between the SSM/I and conventional (ship and island) wind observations. The aircraft reconnaissance winds obtained for 2110 UTC 13 July 1987 were also added to this sample.

Table 3: showing the maptimes for which the SSM/I and (+ or -3h) conventional winds were compared.

TIME	TROPICAL CYCLONE
2110 UTC 13 July 1987	Typhoon Thelma
0940 UTC 14 July 1987	" "
2200 UTC 24 October 1987	Typhoon Lynn
1035 UTC 25 October 1987	" "
2150 UTC 25 October 1987	" "
1023 UTC 26 October 1987	" "

Fig. 12 shows the correlation coefficient and regression estimates between the SSM/I winds and all winds (ship and island winds for map-times in Table 3 and reconnaissance winds numbering 22 for 2110 UTC 13 July 1987). Since satellite winds are representative of an area of 850 sq. km the conventional winds in such an area (where SSM/I winds existed) are averaged and compared. A good correlation of 0.73 resulted. This may be compared to a coefficient of 0.50 obtained by Holliday and Waters (1988). Their sample (numbering 14) consisted of tropical buoydata for January - April 1988 and the correlation coefficient has to be reevaluated for a larger data set.

From the regression line in Fig. 12 it is also clear that the SSM/I overestimated in the lower range up to 15 m s^{-1} and underestimated above 22 m s^{-1}. These findings are somewhat similar to those of Black et al. (1986) who found a bias in the SEASAT-A Scatterometer System (SASS). The SASS was an active microwave instrument (14.6 GHz) having a resolution of 50 km. The bias was higher for low winds and low for winds exceeding 10 m s^{-1}.

Fig. 13 shows a nonlinear fit to the same data that were used in Fig. 12. A slight improvement in the correlation was registered.

The correlation was examined between the SSM/I winds and ship winds for the time periods in Table 3. The sample size (N) was 188. A high correlation of 0.73 was obtained (Fig. 14). This shows that over the open ocean the SSM/I winds are fairly representative. One cannot say this for land winds (in the vicinity of islands and the coast) because the correlation dropped to 0.67 (Fig. 15).

6. Conclusion

The results discussed above had shown that the SSM/I produced precipitation patterns in typhoons which are similar to those obtained by a land based radar. The resolution of rain bands by SSM/I was somewhat poorer compared to the radar. This happens to be the case because the 37 GHz channel of the SSM/I forms a footprint which has a (course) resolution of approximately 40 km. Good correlation is indicated in the medium range in the vicinity of 5 mm h^{-1}.

The SSM/I winds and conventional winds were also compared for about five maptimes involving two typhoons. The typhoon center was enveloped by a 30 m s^{-1}. Although higher wind speeds existed near the eye they were not so discerned by the SSM/I. Winds in the medium range between 15 to 22 m s^{-1} appear to be well represented by the SSM/I.

Higher resolution of wind and precipitation patterns is feasible if the 85 GHz channel brightness temperatures are utilized. Plans are underway to use these brightness temperatures and improve further the correlation coefficients between the SSM/I measured and conventional

parameters. Similarly additional research is also necessary to further exploit these brightness temperatures for the short term intensity changes of typhoons.

Acknowledgments

This research is sponsored by the Air Force Office of Scientific Research under Contract Number F49620-85-C-0013 through the Universal Energy System. I received valuable help in the data analysis and computations from Edward Ciardi, Captain, U.S. Air Force. Various suggestions in the performance of this research were also made by Dr. Kenneth R. Hardy and Mr. Morton Glass of the Air Force Geophysics Laboratory. Mr. Gerald W. Felde of the same Laboratory also generously supplied the SSM/I data. The radar data were supplied by Mr. Tsung Yao Wu, Director of the Central Weather Bureau and his colleagues Dr. Tai-chi Wang of the National Central University and Mr. Lai-Fa Chen. Mr. Keith Nieman of the Computing and Information Systems of St. Louis University assisted us in the computations.

REFERENCES

Battan, L. J., 1973: *Radar Observation of the Atmosphere*. University of Chicago Press, 324 pp.

Black, P. G., R. C. Gentry, V. J. Cardone and J. D. Hawkins, 1986: SEASAT microwave wind and rain observations in severe tropical and mid-latitude marine storms. Advances in Geophysics, 21, 197-277.

Bunting, J. T., and K. R. Hardy, 1984: Cloud identification and characterization from satellites, in a Satellite Sensing of a Cloudy Atmosphere: Observing the Third Planet, Edited by A. Henderson. Sellers, 203-240, Taylor and Francis, London.

Holliday, C. and K. Waters, 1988: SSM/I wind alogrithm evaluation; late winter-spring buoydata. Unpublished manuscript. AFGWC/WFMP, 8 pp. Available from Col. R. P. Wright, AFGWC, Offutt AFB, Nebraska.

Hollinger, J., and R. Lo, G. Poe, R. Savage and J. Pierce, 1987: Special Sensor Microwave/Imagery, *User's Guide*. Naval Research Laboratory, Washington, DC.

Hsu, S. a., 1986: Correction of land based wind data for offshore applications. A further evaluation. *J. Phys. Ocean.*, *16*, 390-394.

Jorgensen, D. P., and P. T. Willis, 1982: A Z-R relationship for hurricanes. *J. Appd. Meteor.*, *21*, 356-366.

Liou, K.N., 1980: *An Introduction to Atmospheric Radiation*. Academic Press, 392 pp.

Lovejoy, S. and G. L. Austin, 1980: The estimation of rain from satellite - borne microwave radiometers. *Q. J. Roy. Met. Soc.*, *106*, 255-276.

Raschke, E. and R. Ruprecht, 1981: Microwave radiometry sampling problems demonstrated with NIMBUS 5 rainrates versus GATE data. *Precipitation Measurements from Space - Workshop reprint*, ed. by D. Atlas and O. Thiele. NASA Lab. For Atmos. Sci., D84-D93 pp.

Rao, M.S.V., 1984: Retrieval of world wide precipitation and allied parameters from satellite microwave observations. Advances in Geophysics, 26, 237-336.

Rogers, R. R., 1976: *A Short Course in Cloud Physics*. Pergamon Press, 227 pp.

Spencer, R. W., H. M. Goodman, and R. E. Hood, 1988: Precipitation retrieval over land and ocean with the SSM/I: Identification and characteristics of the Scattering Signal. Accepted by J. Atmos. and Oceanic Tech.

Staff, Joint Typhoon Warming Center, 1987: Annual Tropical Cyclone Report. NAVOCEAN COMCEN/JTWC, 213 pp.

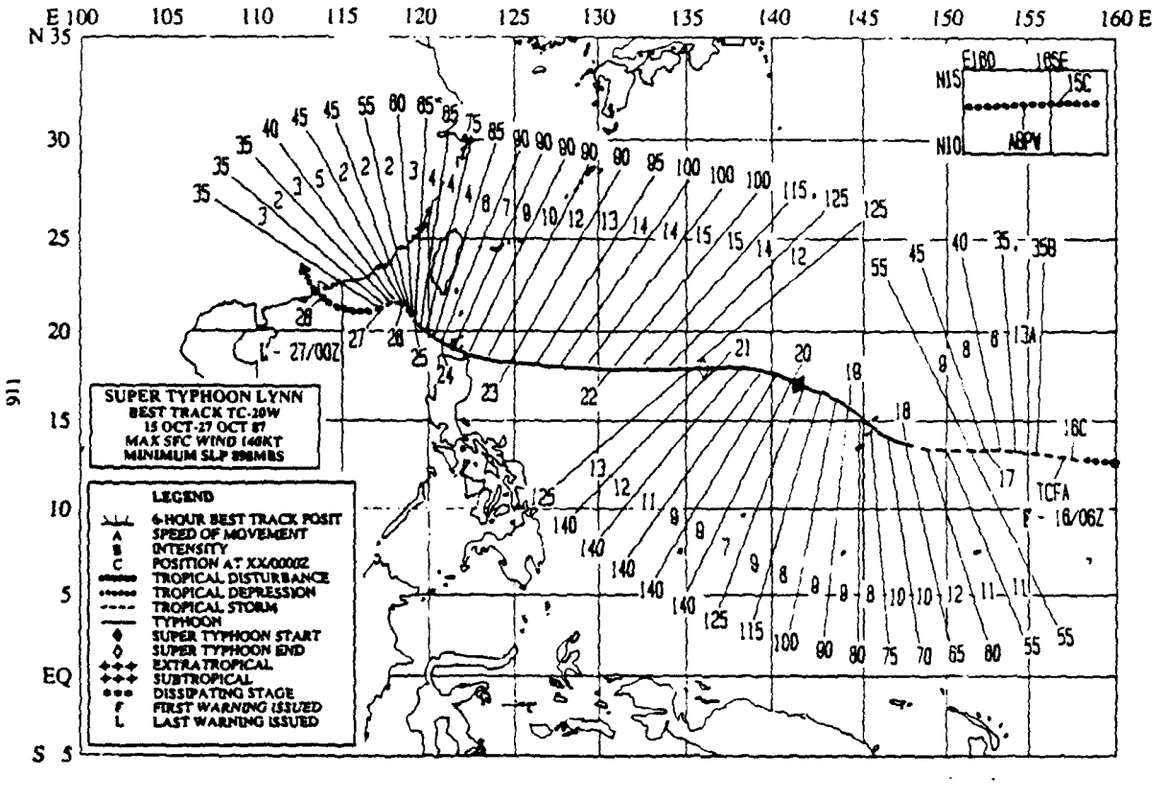


Fig. 1. Track of typhoon Lynn. Of interest in this study is the period October 24-26, 1987 when typhoon Lynn was close to Taiwan. (Staff, JTWC, 1987).

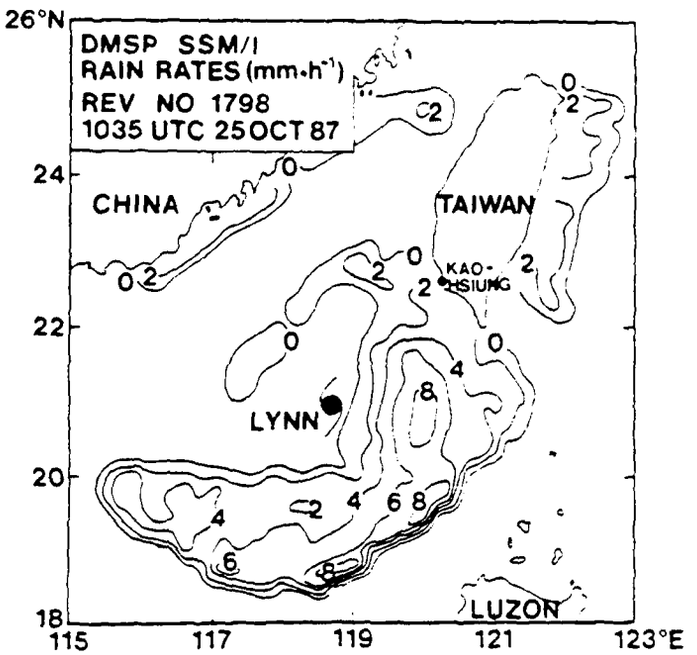


Fig. 2. The DMSP SSM/I rainrates (mm h^{-1}) for 1035 UTC 25 October 1987. Notice the extreme rainrates in south east of the typhoon center.

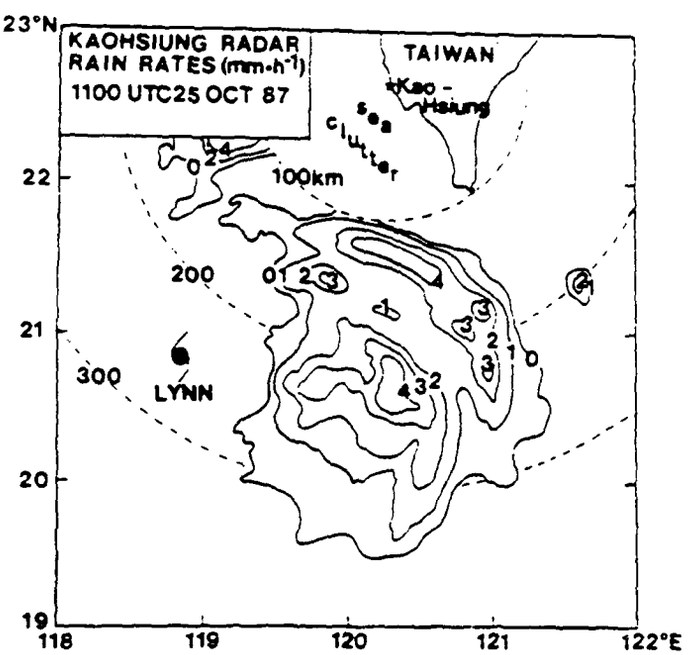


Fig. 3. Rainrates derived under the Jorgensen and Willis (1982) formulation from the reflectivity values of the Kaohsiung radar. Note the superior resolution of the rainbands.

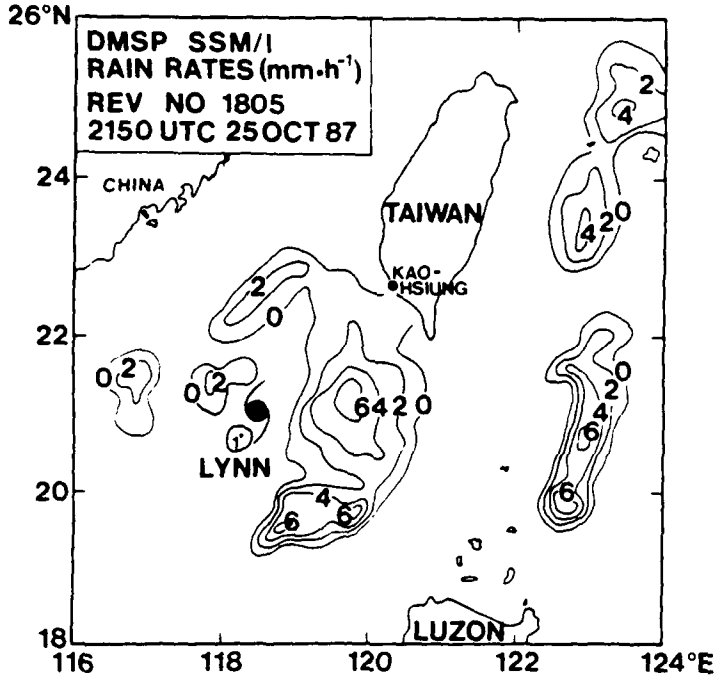


Fig. 4. As in Fig. 2. Note the slow movement of the storm and the shrinkage of the area of precipitation.

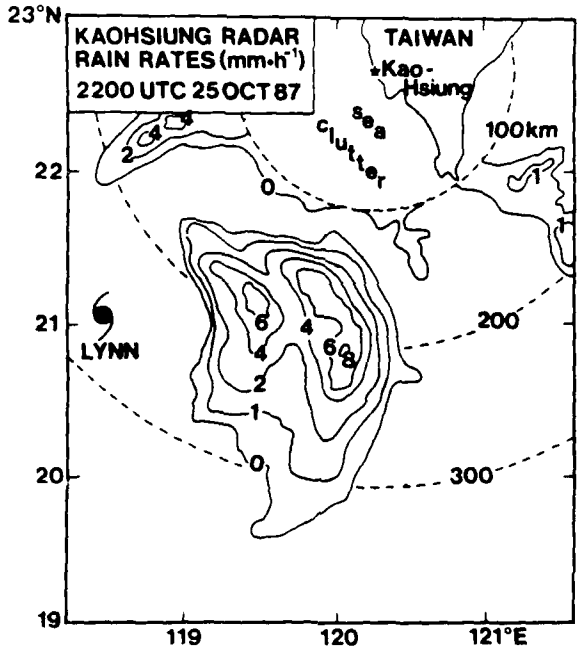


Fig. 5. As in Fig. 3. Of interest are the two rainbands just east of typhoon Lynn.

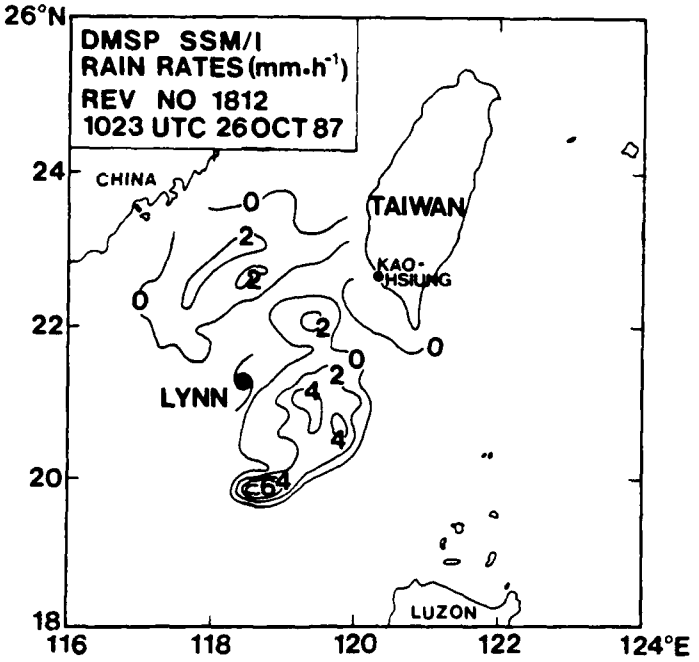


Fig. 6. As in Fig. 2. Note the general reduction of rainrates.

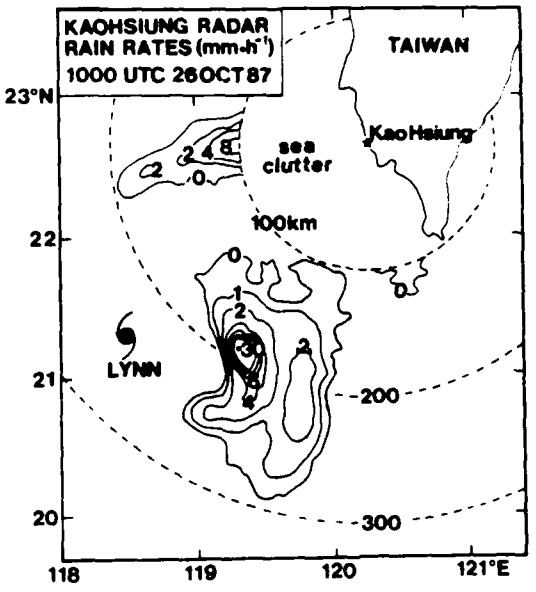


Fig. 7. As in Fig. 3. Note the rainrate extremum east of the storm center.

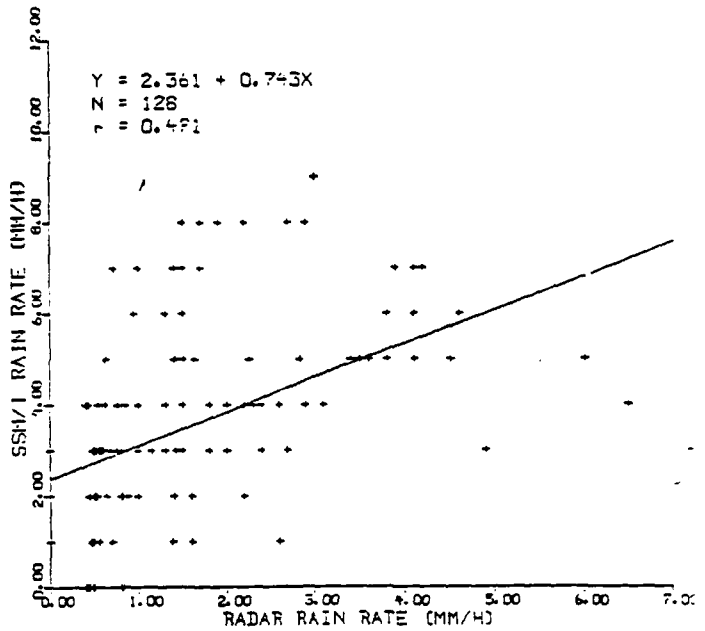


Fig. 8. Regression equation and correlation coefficient (r = 0.49) between the SSM/I rainrates and radar rainrates under the JW formulation.

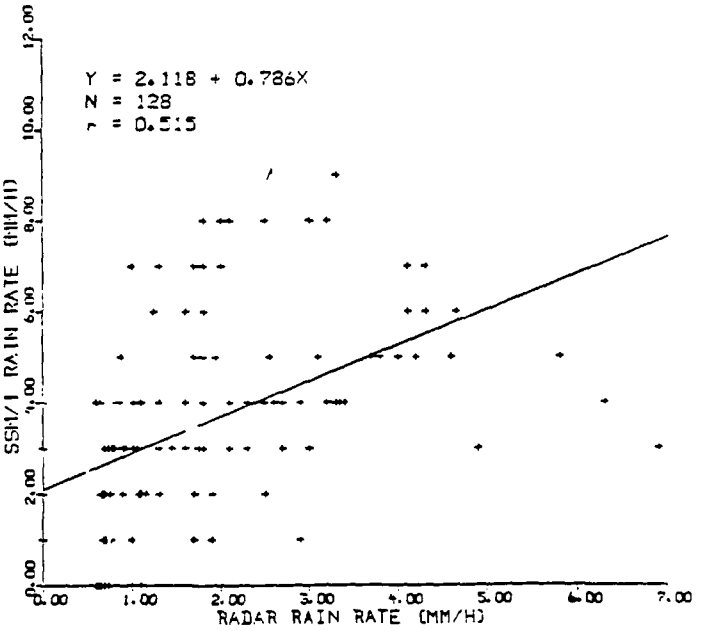


Fig. 9. As in Fig. 8 except the rainrates were derived under an MP formulation.

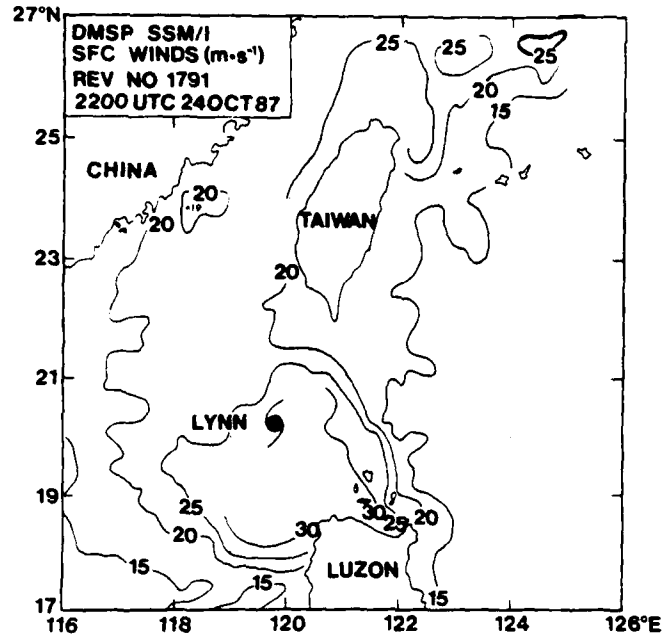


Fig. 10. The DMSP SSM/I winds. Note the 30 m s^{-1} envelope surrounding the center.

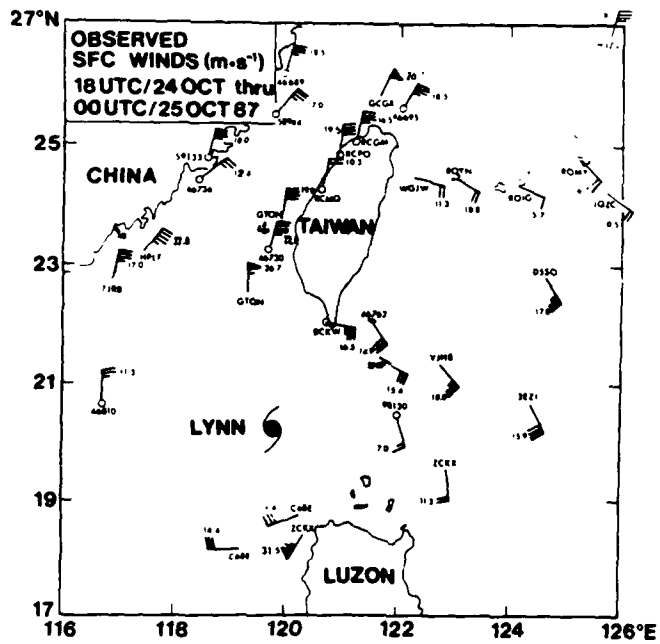


Fig. 11. Sea level winds + or - 3 h centered on 22 UTC (SSM/I time) 24 October 1987.

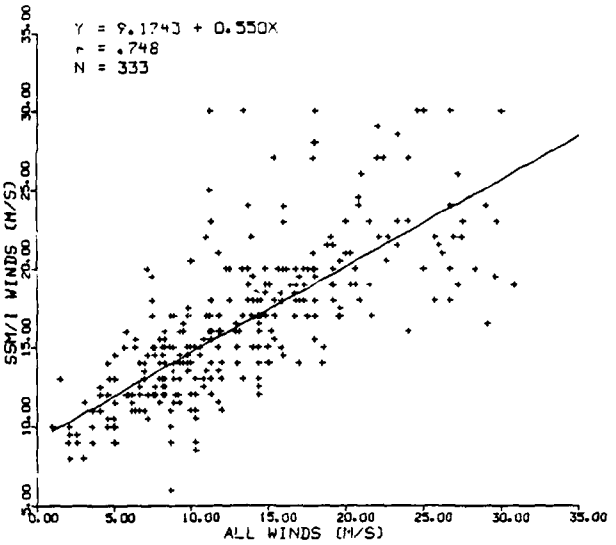


Fig. 12. A linear regression analysis of the SSM/I and all winds (ship, coastal, island and reconnaissance).

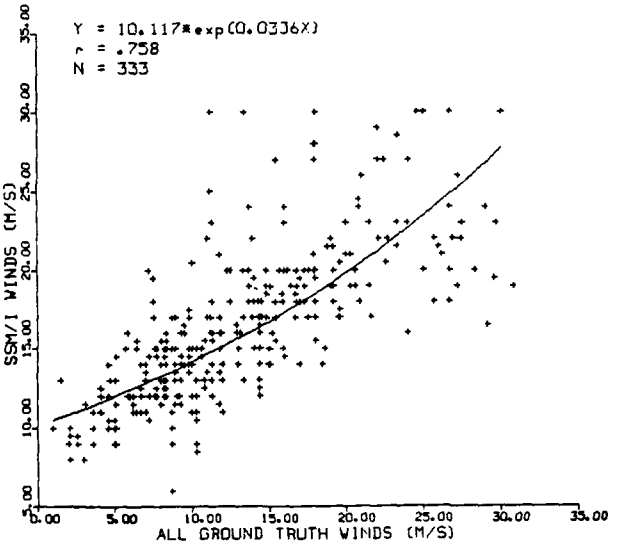


Fig. 13. As in Fig. 12 except a nonlinear fit is indicated.

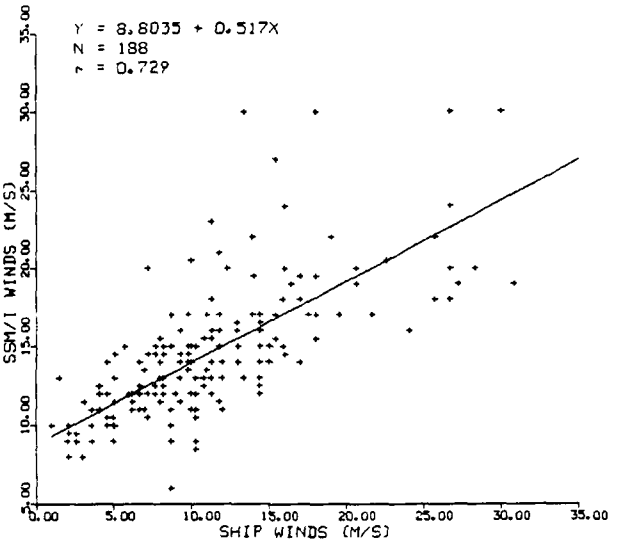


Fig. 14. As in Fig. 12 except the correlation is between the ship winds and the corresponding SSM/I winds.

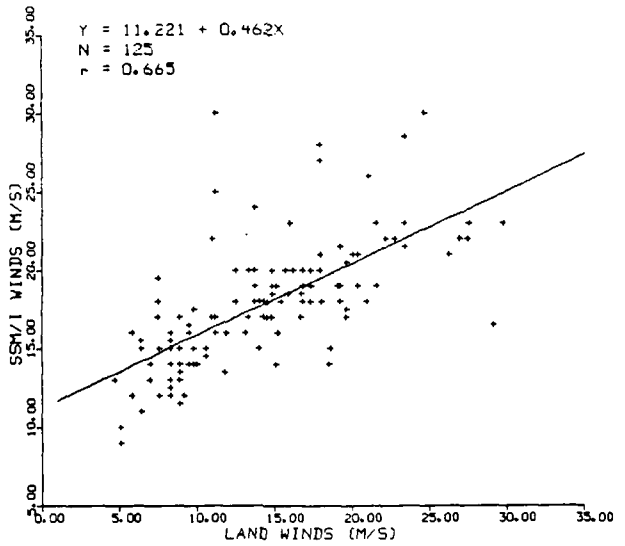


Fig. 15. As in Fig. 12 except the correlation is between the land (coastal and island) winds and the corresponding SSM/I winds.

FINAL REPORT NUMBER 36
REPORT NOT AVAILABLE AT THIS TIME
Dr. Timothy Su
760-7MG-040

1988 USAF-UES RESEARCH INITIATION PROGRAM GRANT

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

FINAL REPORT

DEVELOPMENT OF A SYSTEM FOR THE MEASUREMENT OF ELECTRON EXCITATION CROSS SECTIONS OF ATOMS AND MOLECULES IN THE NEAR INFRARED

Prepared by:	Dr. Keith G. Walker
Department & University:	Physics Dept. Pt. Loma Nazarene College
Location:	3900 Lomaland Drive San Diego, CA 92106
Date:	February 21, 1989
Contract No:	S-760-7MG-074

DEVELOPMENT OF A SYSTEM FOR THE MEASUREMENT OF ELECTRON EXCITATION CROSS SECTIONS OF ATOMS AND MOLECULES IN THE NEAR INFRARED

by

Keith G. Walker

ABSTRACT

A system has been designed and constructed for the obtainment of optical electron excitation functions for wavelengths out to 2650 nm. To produce such a system, much attention was given to reducing unwanted background signals. A collision chamber for the observation of optical emission resulting from electron-atom collisions has been designed so that a minimal amount of scattered radiation from background sources is presented to the viewing detector. A comparison has been made between observation of emission lines in helium with a warm line filter vs. a cold line filter. The detection capability of the constructed system has been evaluated when cold line filters are the mechanism of line isolation. A monochromator and PbS detector have been coupled together. This assembly has been modified so as to minimize noise due to thermal radiation. The resultant detection system has then been utilized to observe optical excitation functions for the 1083 nm and 2058 nm lines of helium from threshhold to 400 eV. The detection capability of the system has been evaluated when line isolation is accomplished with the monochromator. Optical excitation functions of xenon were then measured for transitions producing wavelengths out to 2650 nm.

INTRODUCTION

Considerable effort has been directed towards theoretical and experimental investigation of electron impact cross sections of atoms and molecules. For the experimentalist this has meant, if using optical techniques, studying gases and their transitions lying in the visible or UV portion of the spectrum. Much could be learned from transitions which give rise to near infrared spectra but experimental investigation in this spectral region is non-trivial and little work has been done. In fact, several areas of interest require such information. The infrared window of the terrestrial atmosphere dictates a continuing interest in development of lasers of a variety of wavelengths and power levels in the near IR. Much effort is also being directed into the study of mechanisms that give rise to atmospheric infrared emission phenomena and to gain insight into these mechanisms one must observe the effects of electron-molecule interactions upon the infrared emission spectra. The LSI branch of the AF Geophysics Lab at Hanscom AF Base is one laboratory that is involved in such studies. To obtain data from electron-molecule interactions at low number densities in the infrared region it developed the unique LABCEDE facility. The LABCEDE facility consists of a large cylindrical vacuum tank, 3.4 m long and 1 m in diameter, in which an electron beam is injected. Inside the outer chamber a cylindrical shroud exists which is cooled to liquid nitrogen temperatures. The engineering of this 'dual walled' chamber has been done so efficiently that when cooled the

background radiation as seen by their detectors is equivalent to 88 ^{o} Kelvin. Presently, the system is so designed that only high voltage electron beams are available (2 keV to 5 keV). Such beam energies are currently adequate for the types of information they are studying and produces (with the associated high current densities) sufficiently large signal strengths for detection.

The largeness of the LABCEDE facility is necessary in order to perform a greater number of different kinds of experiments. For example, numerous kinetic processes can be perturbed by 'wall effects' and having a flowing gas system with large dimensions enables accurate data obtainment. However, there are many situations where the large chamber is not necessary. It would be valuable to have a "mini" LABCEDE for the observation of near infrared radiation arising from electron interaction with atoms and molecules. Hence, this Research Initiation Project is for the purpose of beginning the initial phases of research on the construction and utilization of a small system for the observation of radiation from 1000 nm to 2500 nm arising from electron impact upon atoms and molecules. Not only would such a system be smaller and less expensive to operate for those experiments where wall effects are not important but would offer electron beam energies and energy resolutions that are presently not available on LABCEDE. It is herein the goal to develop a system for the expressed purpose of measuring electron impact excitation cross sections in the near infrared spectral region. Such a system would be unique and valuable.

OVERVIEW OF RESEACH EFFORT

The first effort of this research focused on the problem of creating an interaction region between electrons and atoms that would present to a PbS detector a minimum of thermal and scattered radiation. Secondly, attention was given to the detector itself so that optimum signal-to-noise ratio (SNR) could be obtained under the conditions presented to the detection system. Finally, spectra and electron-impact excitation functions were to be obtained out to 2500 nm for helium and xenon in order that the capabilities of the system might be evaluated.

All of the above goals were met. A chamber was designed and constructed which was quite effective in eliminating the scattered light while at the same time enhancing the optical signal and maintaining the integrity of the electron beam's energy resolution and geometry. The PbS detector, its associated dewar, and a 27 cm monochromator were modified successfully and an improved SNR resulted. The electron impact optical excitation functions were measured from threshhold to 400 eV for the 1083 nm and 2058 nm transitions in atomic helium. Optical excitation functions for xenon were observed to 400 eV for transitions giving rise to wavelengths as high as 2650 nm.

From this work, an evaluation was made as to the present system's capabilities in measuring electron excitation cross sections and what must be done to improve the system's performance and enhance the output by an order of magnitude or more.

The remainder of this report will detail how each of the above

goals were met and the stated results obtained.

CONSTRUCTION

The initial phase of this research was to modify the present vacuum chamber and viewing region to provide baffling of scattered light from the hot cathode and minimization of thermal radiation from the surrounding environment. Figure 1 is an illustration of the electron gun. This gun will produce a 3 mm beam with an energy resolution of 0.5 eV at 700 microamps. Grid 1 is used to control the current. Its voltage is chopped between 10 volts above cathode (beam on) to 30 volts below cathode (beam off) at a frequency of 200 Hz. Grid 2 is used for focusing and to regulate the current. Its voltage will depend on the gas used. For helium, it will range from 150 to 200 volts above cathode. For xenon the grid 2 voltage will rarely exceed 50 volts above cathode in order to prevent arcing. Arcing was continually a problem with xenon since a large ion density existed due to xenon's huge ionization cross section. Grid 3 is at cathode potential and is called the virtual cathode. This grid removes many secondary electrons produced from the first two grids and enables an improved energy resolution. Grids 4 and 5 are grounded and provide the acceleration from cathode potential. The electrons from grid 5 emerge into a Faraday cup from which the electron-atom collisions are observed. The standard Faraday cup has been replaced with a stainless steel cylinder that extends into the side nipples of the vacuum chamber (Figure 2).

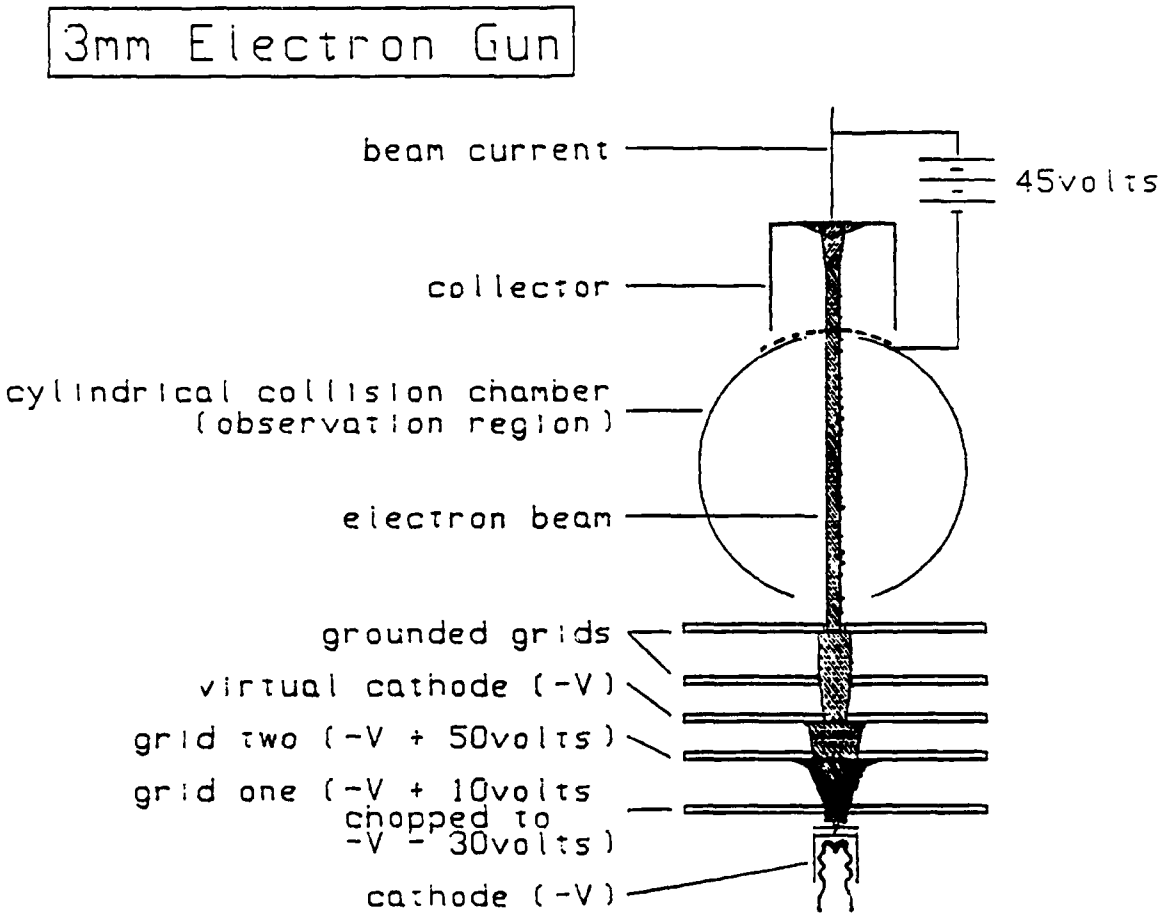


FIGURE 1. 3 mm Electron Gun

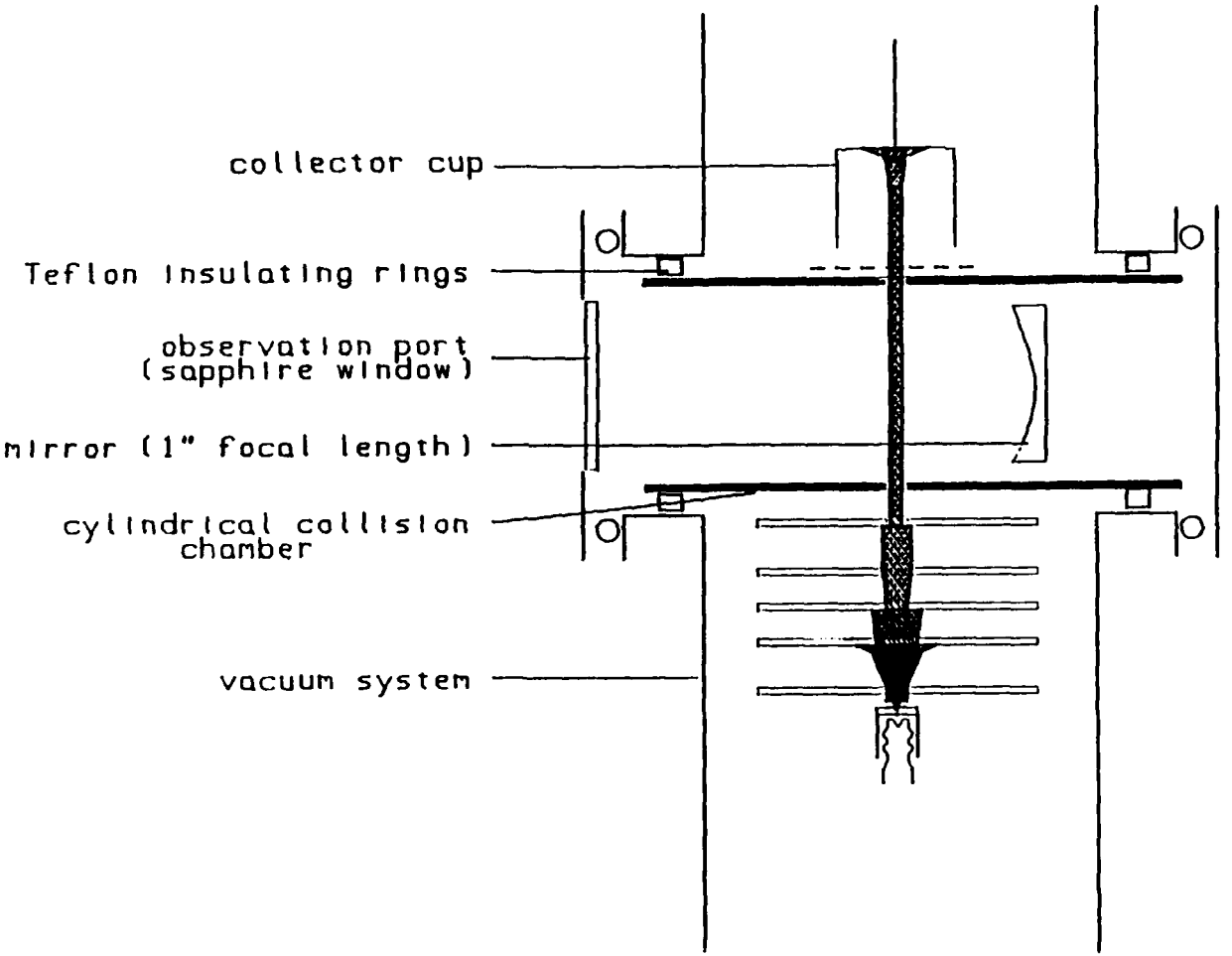


FIGURE 2. Orientation of Cylindrical Collision Chamber in Vacuum System

This configuration provides very effective shielding from the optical emission of the cathode. This also provides a convenient arrangement for the mounting of a 1 inch focal length concave mirror. This aluminized mirror has a MgF_2 coating and is mounted 2 inches from the electron beam thus providing an image of the beam on top of the object beam and an enhancement of signal by a factor of two. The inside of the cylindrical Faraday cup was then blacken with aquadag. The collector cup was biased at +45 volts to suppress secondary electrons from entering the collision region. A wire mesh was inserted over the cylinder's exit hole to prevent field penetration by the +45 volts. The gun is contained in a vacuum which will reach 10^{-8} Torr.

The PbS detector was next given attention. Our first efforts would be to obtain an excitation function of the 2058 nm line in helium ($2\,^1P \rightarrow 2\,^1S$). The most straightforward method of isolating this line is to utilize a line filter. Of particular interest is the difference in SNR between a warm and cold (-196 ^oC) filter measurement. Such a measurement gives an evaluation of the relative importance of cooling the filter. The PbS detector is a 2 mm x 10 mm element mounted on a cold finger of a liquid nitrogen dewar. Capping the dewar viewport is a sapphire window. The dewar with filter is mounted directly to the window of the vacuum system. In the warm filter configuration, the filter was mounted between the chamber window and the dewar's window. In the cold filter configuration, we custom fabricated a filter holder and attached it to the cold finger. This allowed the positioning of the filter directly over

the PbS element. Also, we are able to attach variable sized cold masks over the detector to limit its field of view. Figure 3 illustrates the way the dewar and detector assembly were configured. Figure 4 is the circuitry for the detector. The PbS resistance would reach 20 megohms at liquid nitrogen temperature. If one were to maximize the power delivered to the load resistance then a value equal to the PbS resistance would be choosen for R_L. However, since the electron beam is chopped and lock-in techniques are being utilized, we want to maximize the change in load power due to a change in the resistance of the PbS. This requires that the load resistance be half the PbS resistance (Ref 1). Therefore, the load resistance was set to 10 megohms. Such large resistances and small signals made it necessary to carefully shield the entire dewar, the detector's power supply, and associated leads from stray signals.

In the second phase of the research, a Monospec 27 cm monochromator with f/3.8 speed was used to isolate the lines to be studied. The optical radiation was gathered from the collision region by an f/3, CaF_2 lense. To optimize the SNR, the exit slit housing of the monochromator was removed and the PbS detector and dewar were mounted in its place. The cold mask was made into a slit and positioned so as to become the exit slit of the monochromator. We were able to construct and align this configuration to a degree where the measured bandpasses were within 5% of the regular monochromator exit slit configuration. Also, cold bandpass filters were utilized in front of the slit to cut out the thermal radiation that would lie outside the spectra

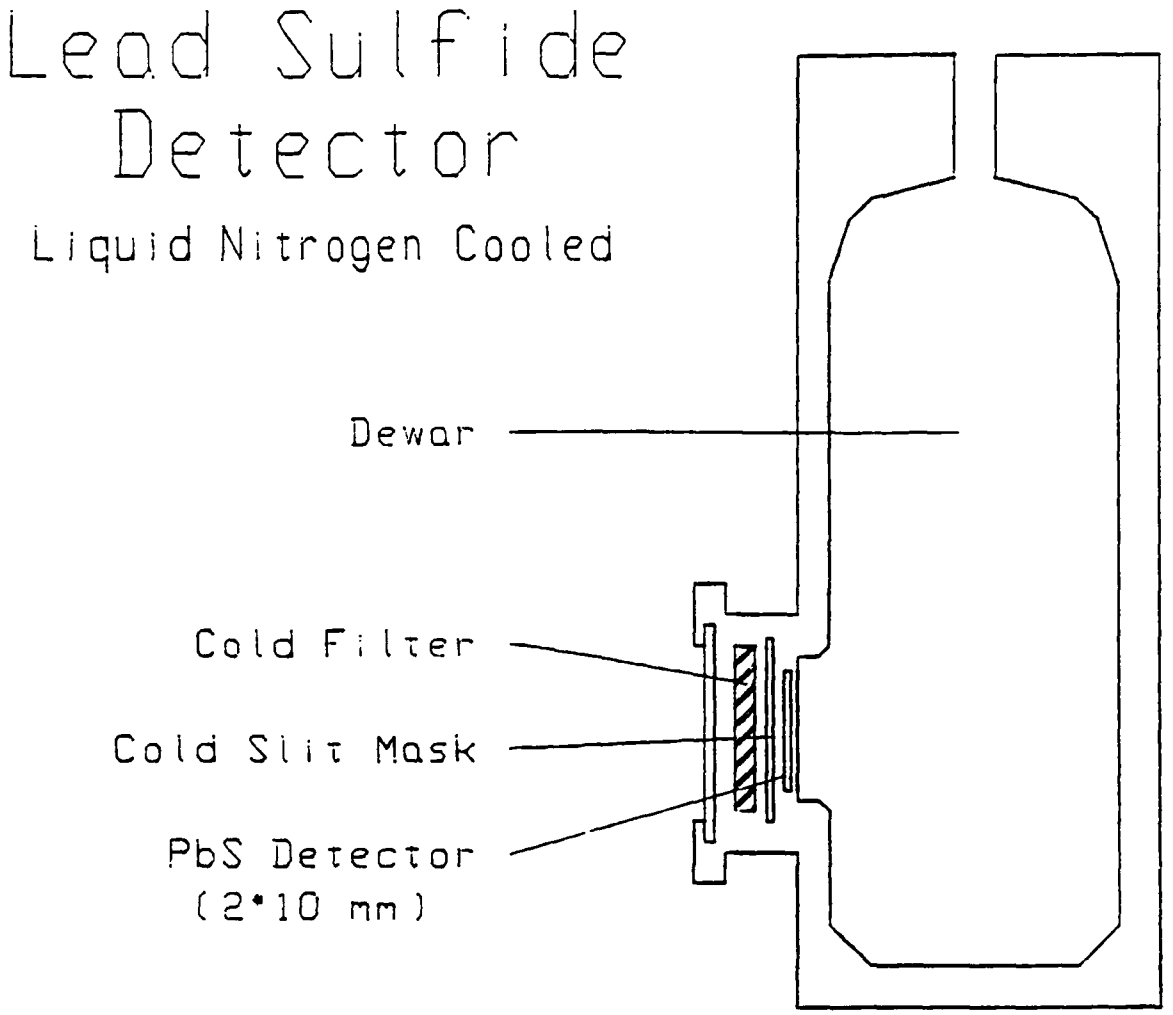


FIGURE 3. PbS Detector and Liquid Nitrogen Dewar

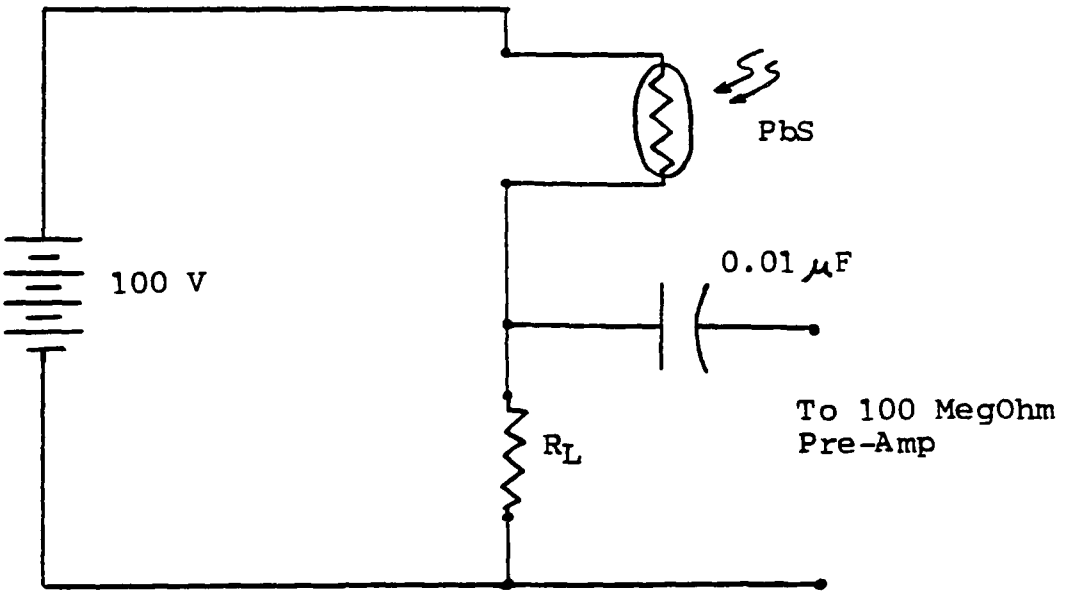


FIGURE 4. PbS Detector Circuit

region we were scanning. These filters covered the following ranges:

* 1000 nm - 1800 nm
* 1800 nm - 3000 nm
* 2000 nm - 3500 nm

Figure 5a and figure 5b show this arrangement. Figure 5a is the normal monochromator exit slit assembly. Figure 5b is the PbS detector and the new cold slit arrangement. This technique of selectively using cold filters in conjunction with the monochromator decreases the noise signal in the region of interest as well as presenting a very cold surface directly in front of the PbS.

Figure 6 is a block diagram of the system. Standard lock-in techniques were used. All data logging and instrument control were performed by a microcomputer. The beam energy was varied; its voltage recorded; the lock-in signal was averaged typically over 20-30 readings. The average detector signal divided by beam current was then plotted against the beam energy in real time on the CRT. Such a plot is designated an excitation function. All data is then stored on disk and hard copy of the excitation function was furnished via digital plotter. Measurement of electron impact excitation functions using such optical techniques is a common method for which a good general description appears in Massey and Burhop (Ref 2).

A later development to the system was the addition of Helmholtz coils for the production of a magnetic field along the electron beam's axis. Fields up to 100 gauss were created and resulted in confining the electrons to a well defined beam.

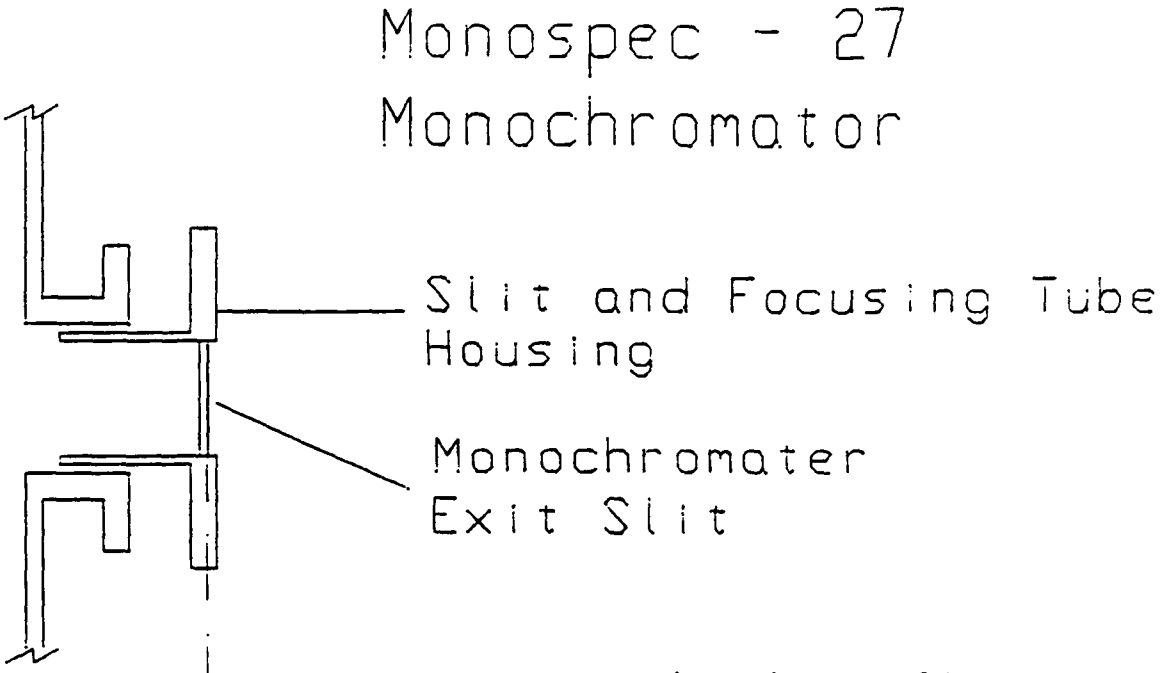


FIGURE 5a. Normal Exit Slit Assembly

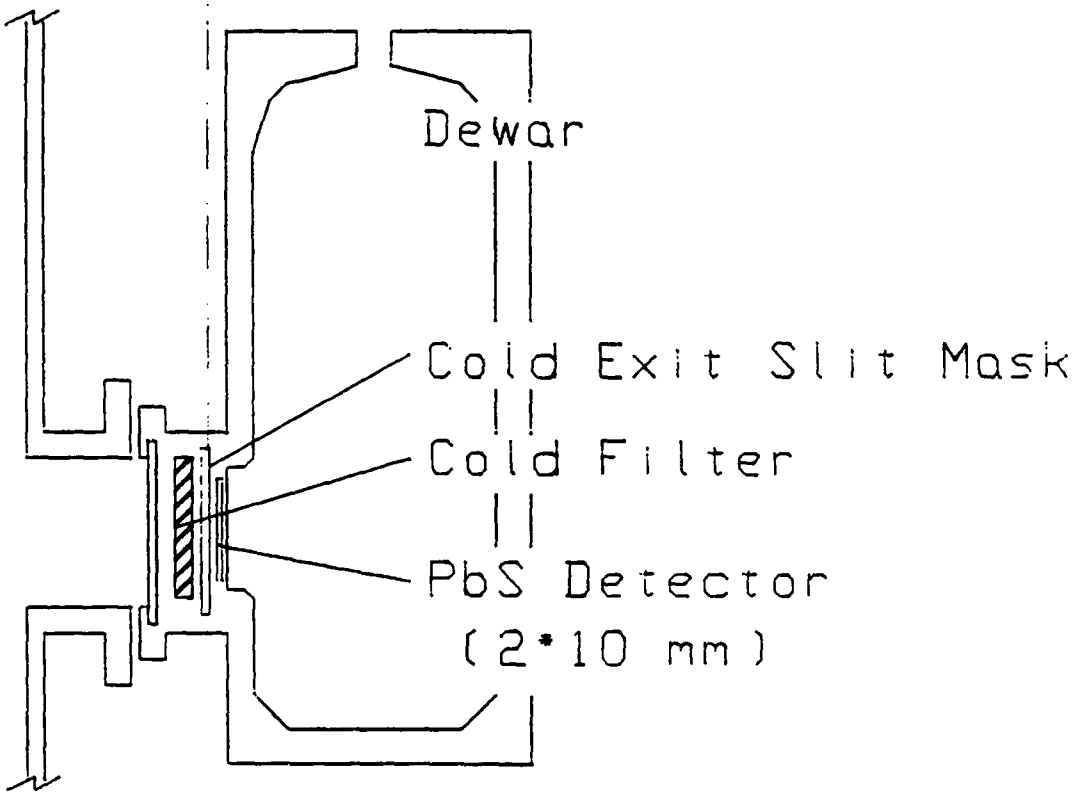


FIGURE 5b. Modified Exit Slit Assembly

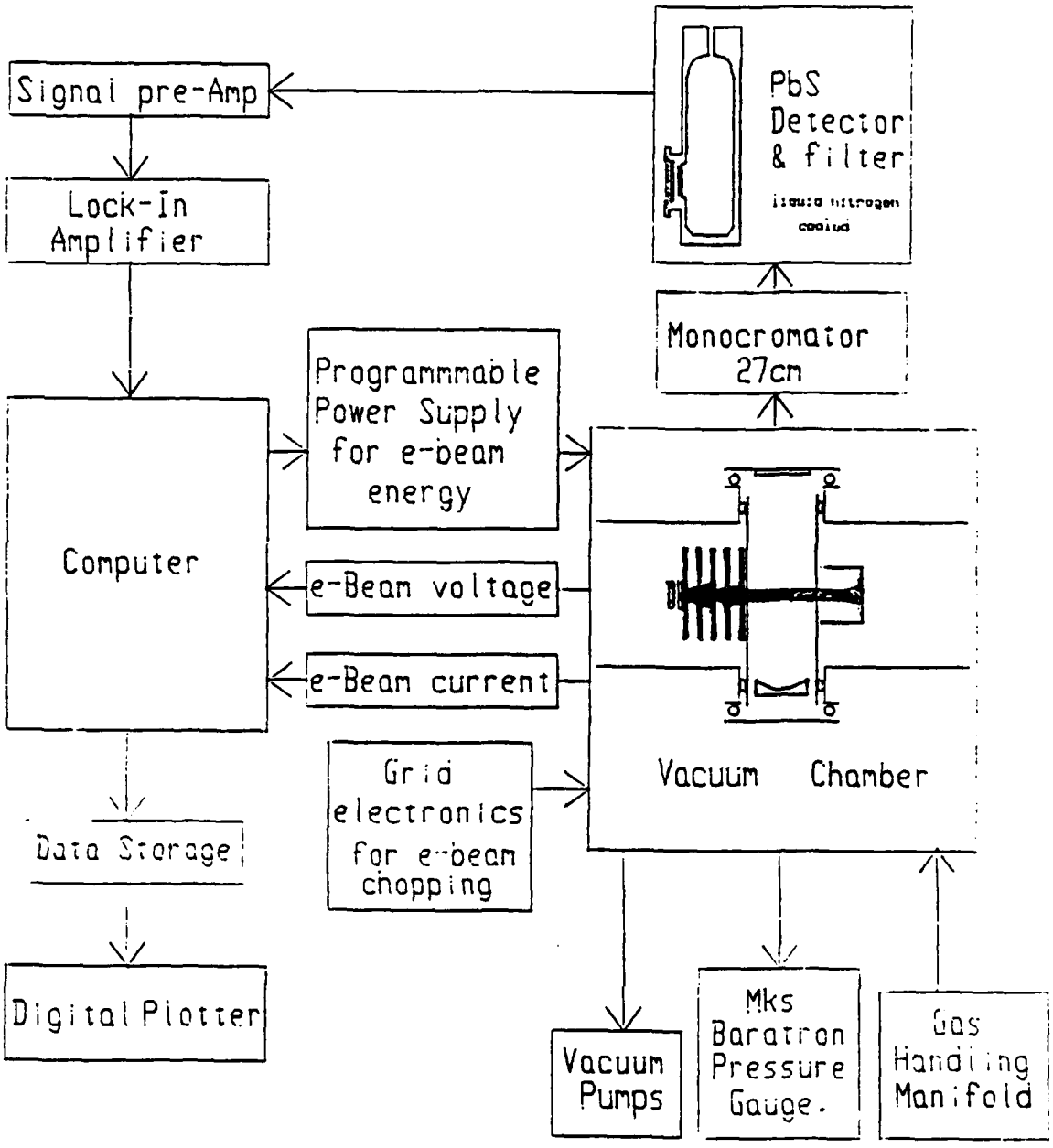


FIGURE 6. Block Diagram of System

The experimental goals were as follows:

1) Measure the 2058 nm and 1083 nm optical excitation function of helium from threshhold to 400 eV using a warm line filter for line isolation and a PbS detector cooled to liquid nitrogen temperatures. Repeat the above measurements using a cold filter (-196 ^{o}C) and compare the differences in SNR.

2) Utilize a 27 cm monochromator in conjunction with cold bandpass filters to observe the 1083 and 2058 nm excitation functions of helium from threshhold to 400 eV. Compare the results with data obtained from goal 1) where line filters were used.

3) Measure optical excitation functions of xenon out to 2500 nm or 3000 nm using the 27 cm monochromator as the spectral dispersion element.

4) Evaluate the capabilities of the system for measurements of emission spectra in the mid-infrared.

The $2\,^{1}P \dashrightarrow 2\,^{1}S$ transition in helium gives rise to the 2058 nm spectral line. This line was choosen because of its isolation, strength and where it lay in the spectrum. The isolation of the line and its location in the spectrum enabled us to utilize a standard off-the-shelf line filter whose band center lay at 2090 nm and whose bandpass was 85 nm. Also, when the filter was put at liquid nitrogen temperatures, the band center dropped to almost the desired wavelength of 2058 nm. The optical cross section of the 2058 nm line is about 1×10^{-19} cm^{2}; however, it is very pressure dependent and when working at high pressures (40

mTorr) an optical excitation cross section of 1×10^{-17} cm^2 should be obtained. This will give us a very strong signal to initiate our measurements. The detector with cold filter were attached to the observation port of the vacuum chamber. This put the detecting element about 5 inches from the collision region. Figure 7 shows the arrangement of system. In this configuration, optical excitation functions were obtained of the 2058 nm line from 40 mTorr to a pressure as low as 1 mTorr. Figure 8 and figure 9 represent the 2058 nm optical excitation function wiith a warm filter and cold filter respectivily. The SNR is significantly better for the cold filter. As was mentioned, this transition is very pressure dependent. Figure 10 is a plot of the emission vs. pressure for an impacting electron of 100 eV.

Since the optical cross section of the $2\,^1P \rightarrow 2\,^1S$ transition is 10^{-19} cm^2, it is evident that by the utilization of line filters one can obtain excitation cross sections in the 10^{-19} cm^2 range. If high pressures can be used without presenting non-linearity effects, 10^{-20} cm^2 would be possible.

The next optical excitation function attempted was the 1083 nm line arising from the $2\,^3P \rightarrow 2\,^3S$ transition. Again a line filter was utilized and the results are given in figure 11. Data was only taken to 200 eV for this case.

The utilization of line filters to resolve spectra allows one to employ large f/values. Unfortunately, line filters are expensive as well as inconvenient and unflexible. Hence, the next phase of this research was to employ a monochromator as the dispersive

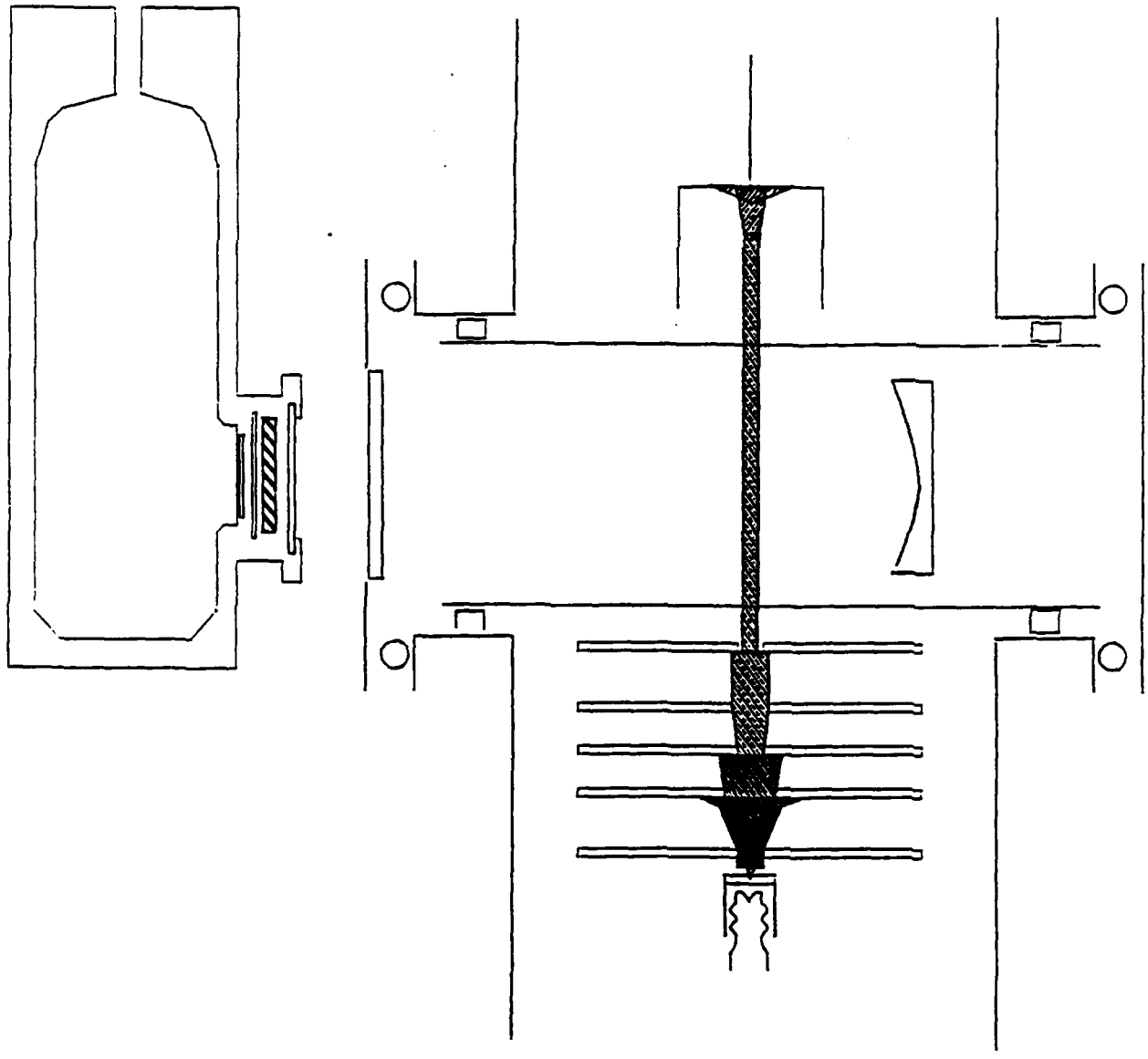


FIGURE 7. Configuration for Detection using Line Filters

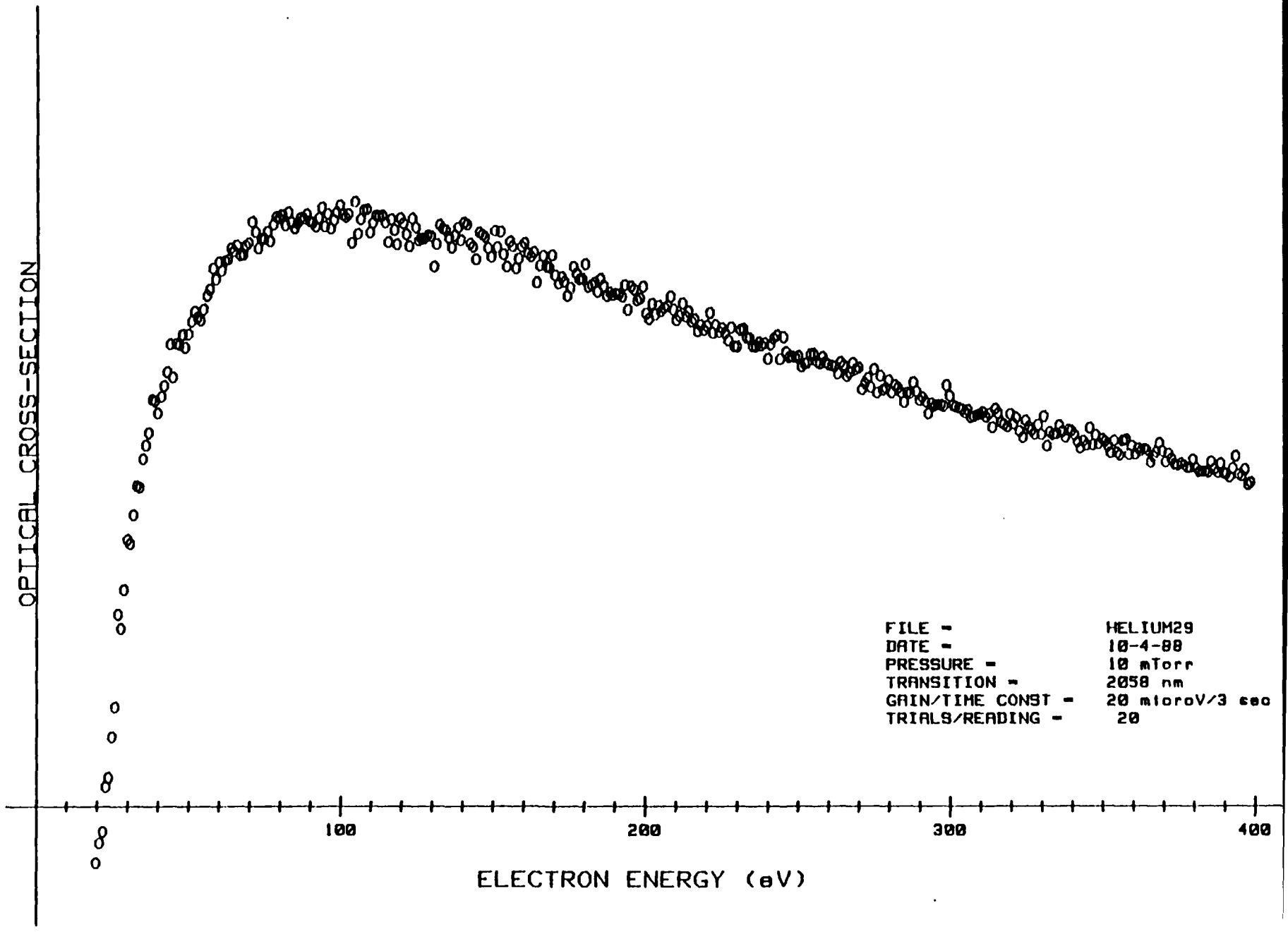


FIGURE 8. The 2058 nm Excitation Function via Warm Line Filter

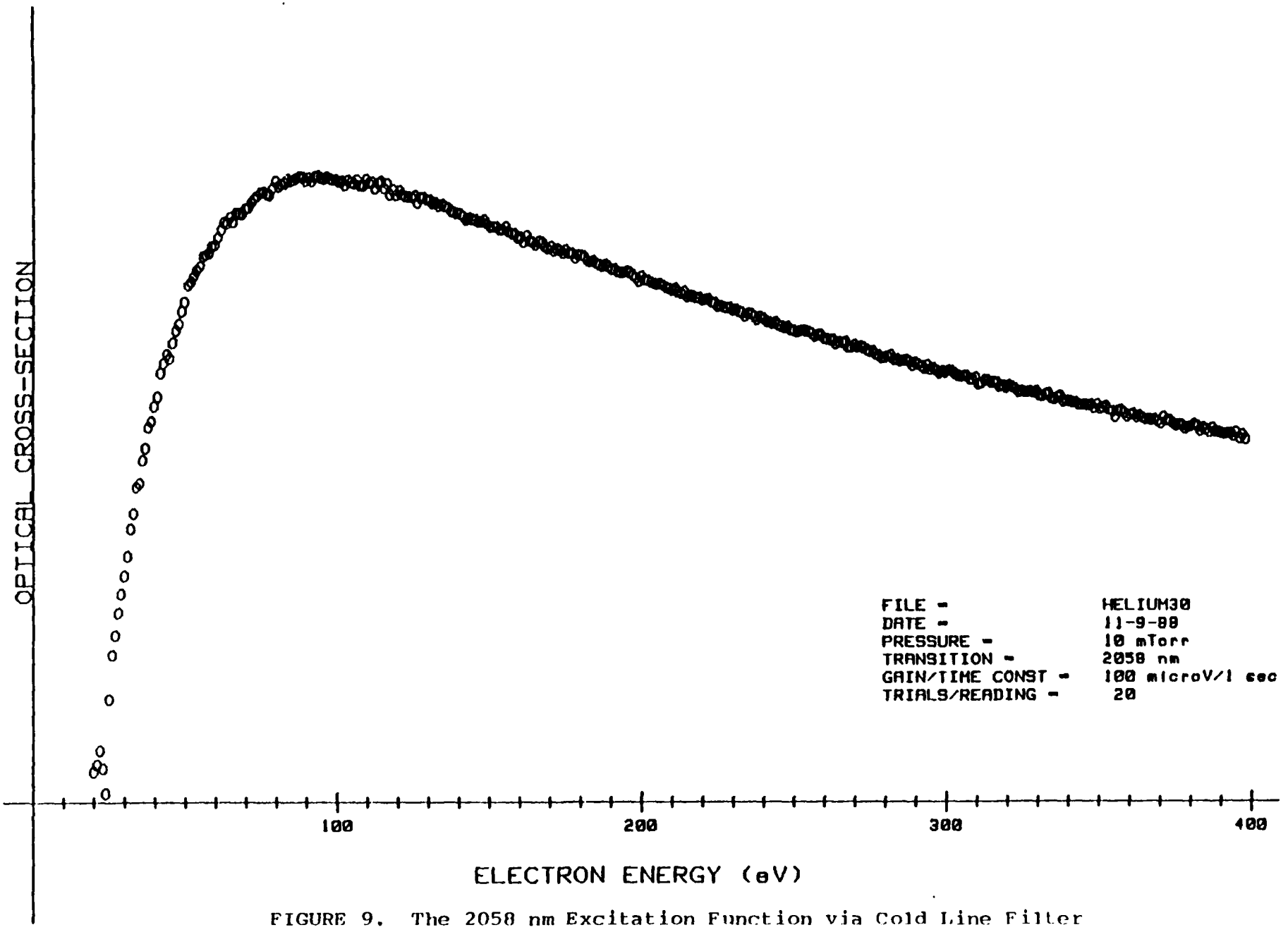


FIGURE 9. The 2058 nm Excitation Function via Cold Line Filter

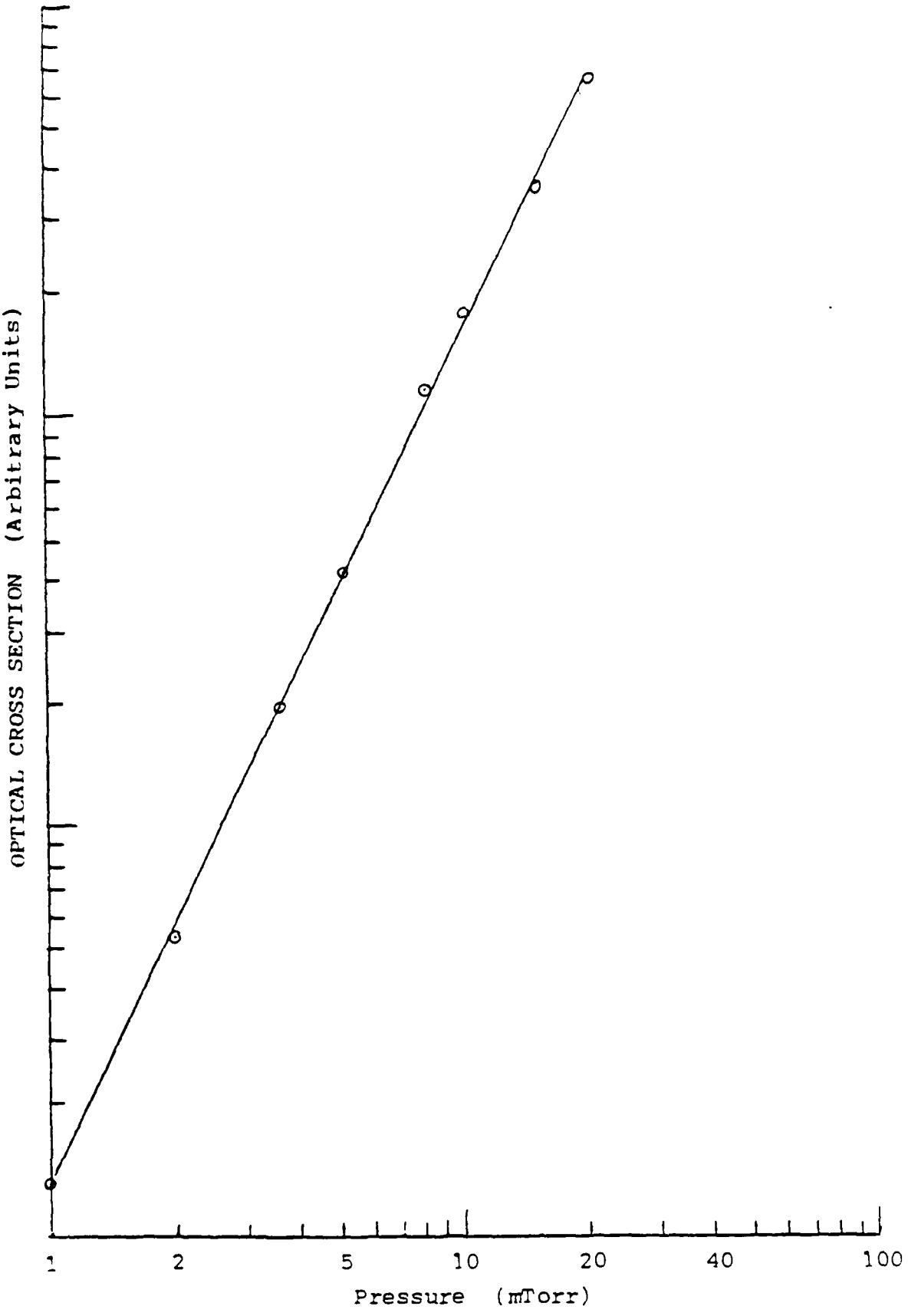


FIGURE 10. Emission vs. Pressure for the 2058 nm line

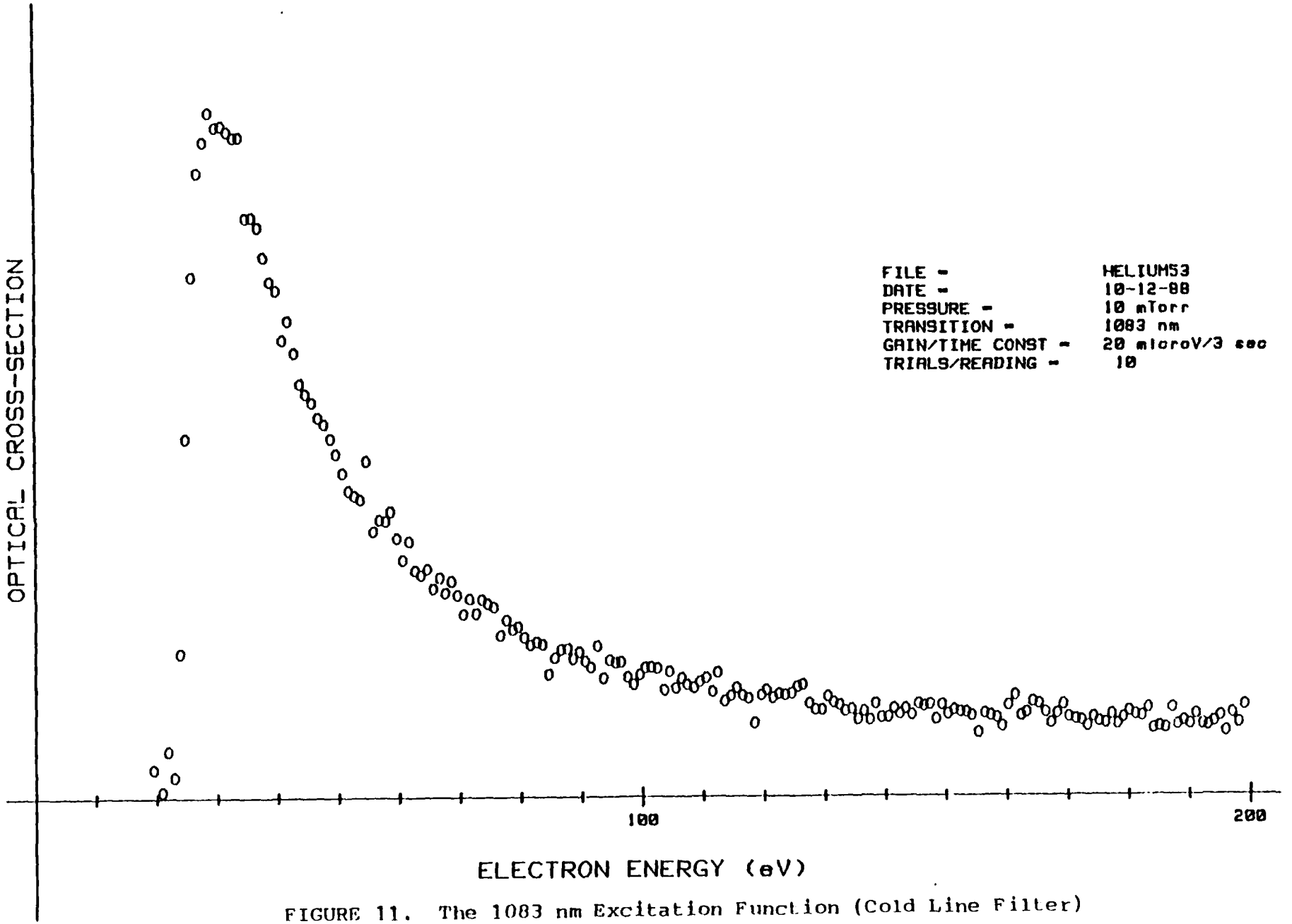


FIGURE 11. The 1083 nm Excitation Function (Cold Line Filter)

medium and evaluate its capabilities.

Figure 5b illustrated the configuration of PbS detector relative to the monochromator. In the normal monochromator exit port, as was shown in figure 5a, a focus tube and slit holder are adjustable for alignment purposes. This assembly was removed and a mount was constructed to attach the dewar and detector onto the monochromator so that a cold copper mask can be positioned where the standard exit slit would normally reside. The copper mask can then be slit to the desired bandpass and it becomes the exit slit for the monochromator. For example, with an entrance and exit slit width of 1 mm a bandpass of 120 Angstroms is expected with the 300 grooves/mm grating blazed at 2000 nm. Our arrangement produced a triangular bandpass with a measured 128 Angstroms --- a 5 % error. Most of this error is within the precision of the slit width construction. Hence, the PbS sees a cold mask directly in front of it with a slit equal to the entrance slit. This also enables us to utilize the entire f/value of the monochromator since the width of the detector element is 2 mm. As mentioned previously, a cold bandpass filter is present to eliminate the unwanted thermal radiation which lays outside the spectral region of interest and which arises from the optical element and/or surroundings that are in the view of the detector.

Figure 12 shows the comparison of excitation functions for the 2058 nm line. The upper curve is taken with the use of a cold line filter; the lower curve with the monochromator (blazed at 2000 nm) and the attached PbS detector. It is evident that the

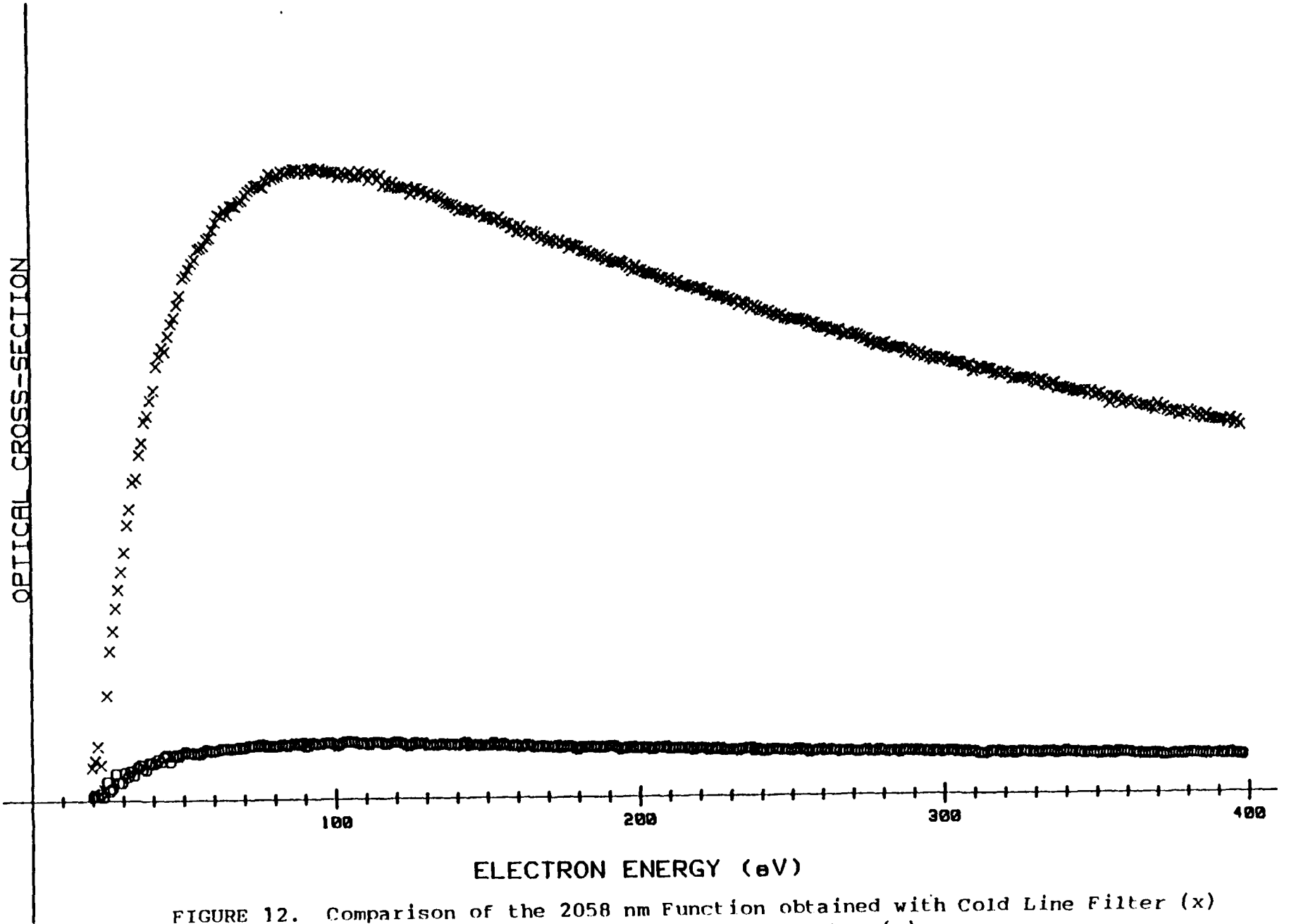


FIGURE 12. Comparison of the 2058 nm Function obtained with Cold Line Filter (x) and that obtained with 27 cm Monochromator (o).

signal strength drops by at least an order of magnitude when using the monochromator. This is due to several reasons:

* Smaller solid angle of observation when using the monochromator.
* Losses occuring in the optics of the monochromator.
* Distance from phenomenon to the detector is significantly farther with the monochromator.

Figure 13 is the same two curves as in figure 12 except we have multiplied the lower curve by a factor of 14 to verify that the shapes of the curves are the same independent of whether the line is isolated by filter or monochromator.

Figure 14 is the $2\,^3P \rightarrow 2\,^3S$ optical excitation function obtained via monochromator with a grating blazed at 1000 nm.

Hence, using the monochromator the present system can obtain excitation cross section of 10^{-18} cm^2 or higher. Again, this could be improved to 10^{-19} cm^2 at high pressures (greater than 10 mTorr).

To utilize this system's detection capabilities we have choosen to examine the 5d --> 6p and 7s --> 6p optical excitation functions. Figure 15 is a partial energy diagram for xenon and it is seen that the 5d and 7s states give rise to a number of spectral lines lying between 1000 nm and 4000 nm. They also have predicted electron excitation cross sections in the 10^{-18} cm^2 range. These transitions have intrinsic interest as well since many of the 5d --> 6p lines are known laser transitions (Ref 3).

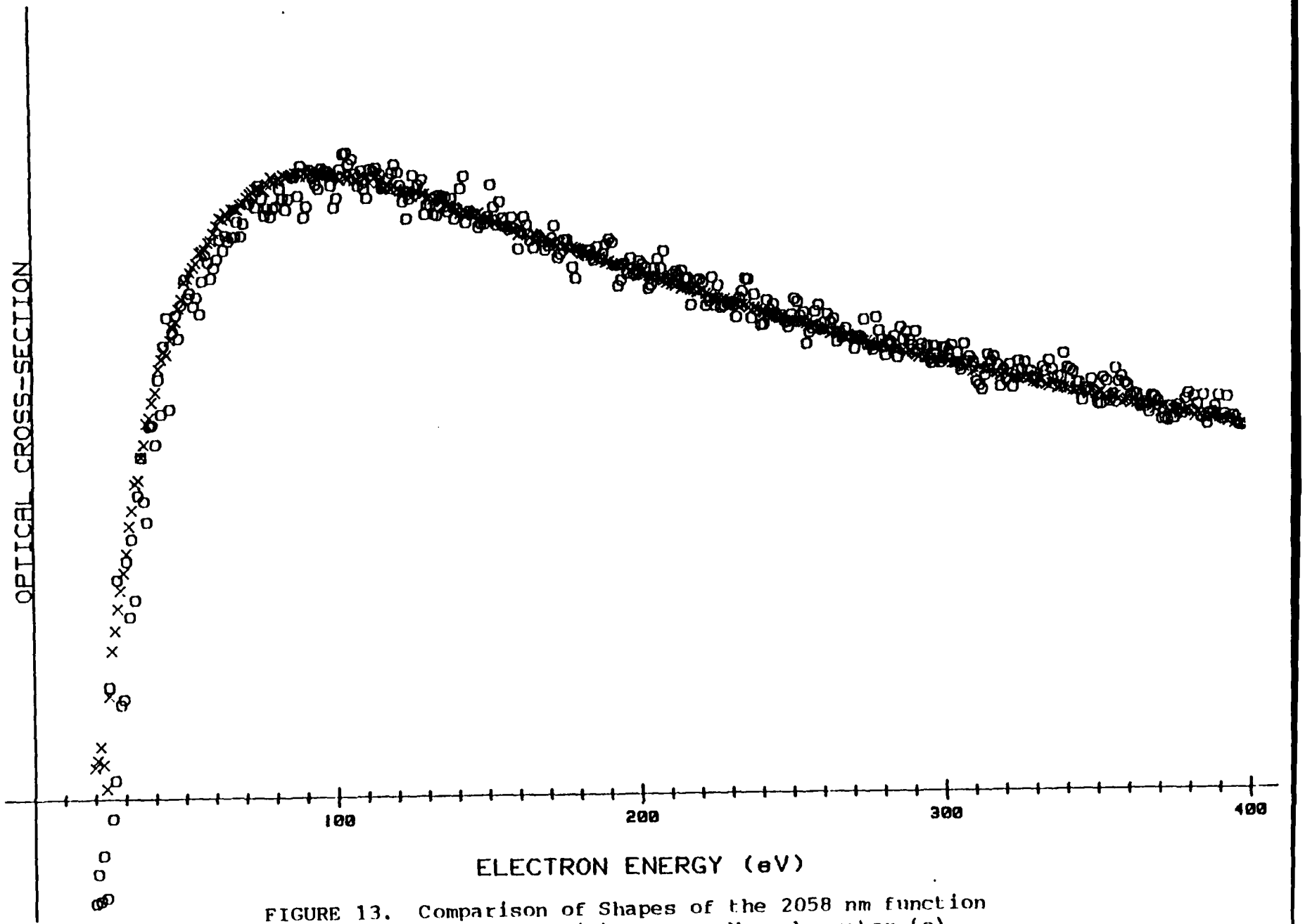


FIGURE 13. Comparison of Shapes of the 2058 nm function
Line Filter (x) Monochromator (o)

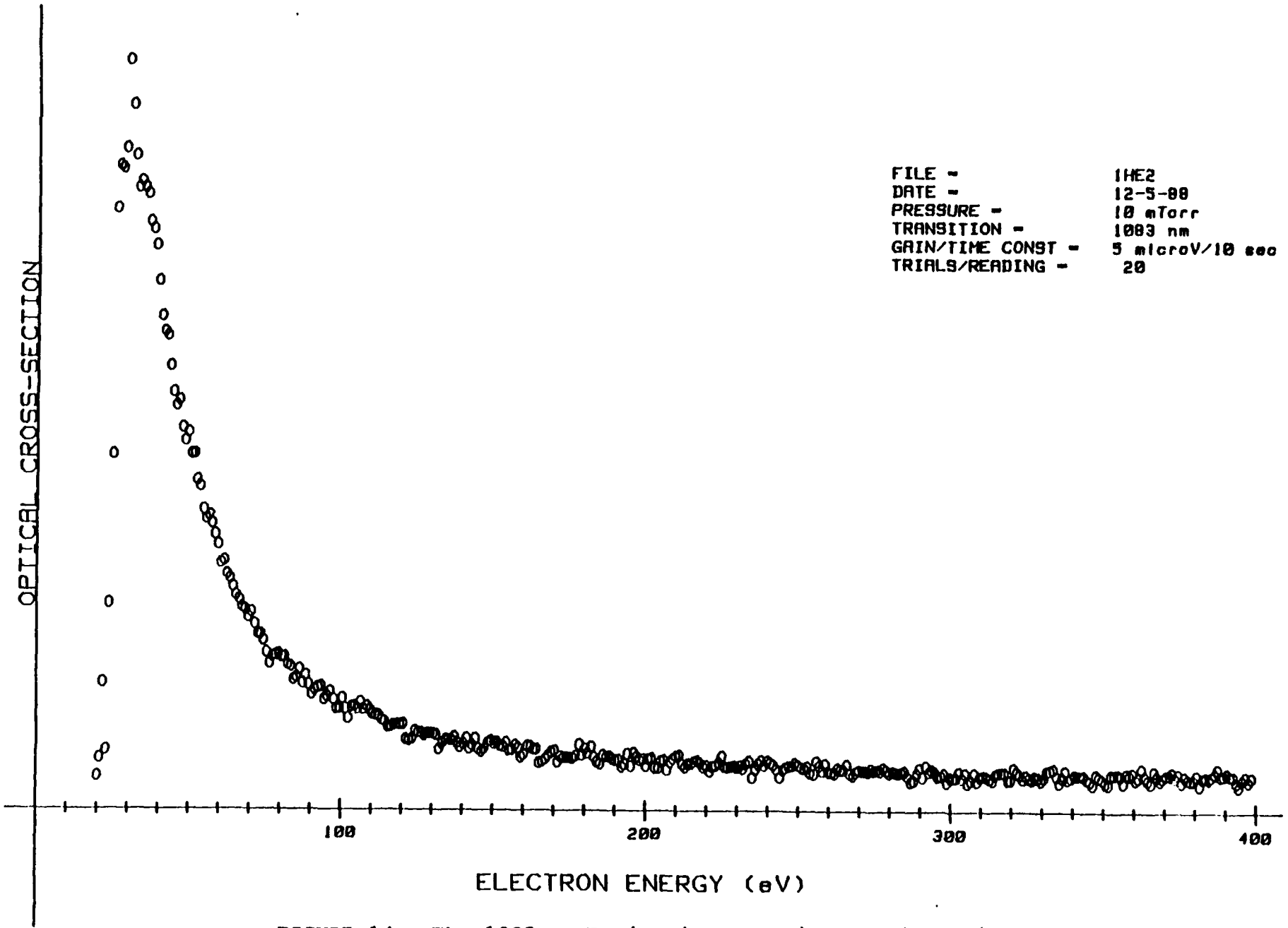


FIGURE 14. The 1083 nm Excitation Function obtained with Monochromator.

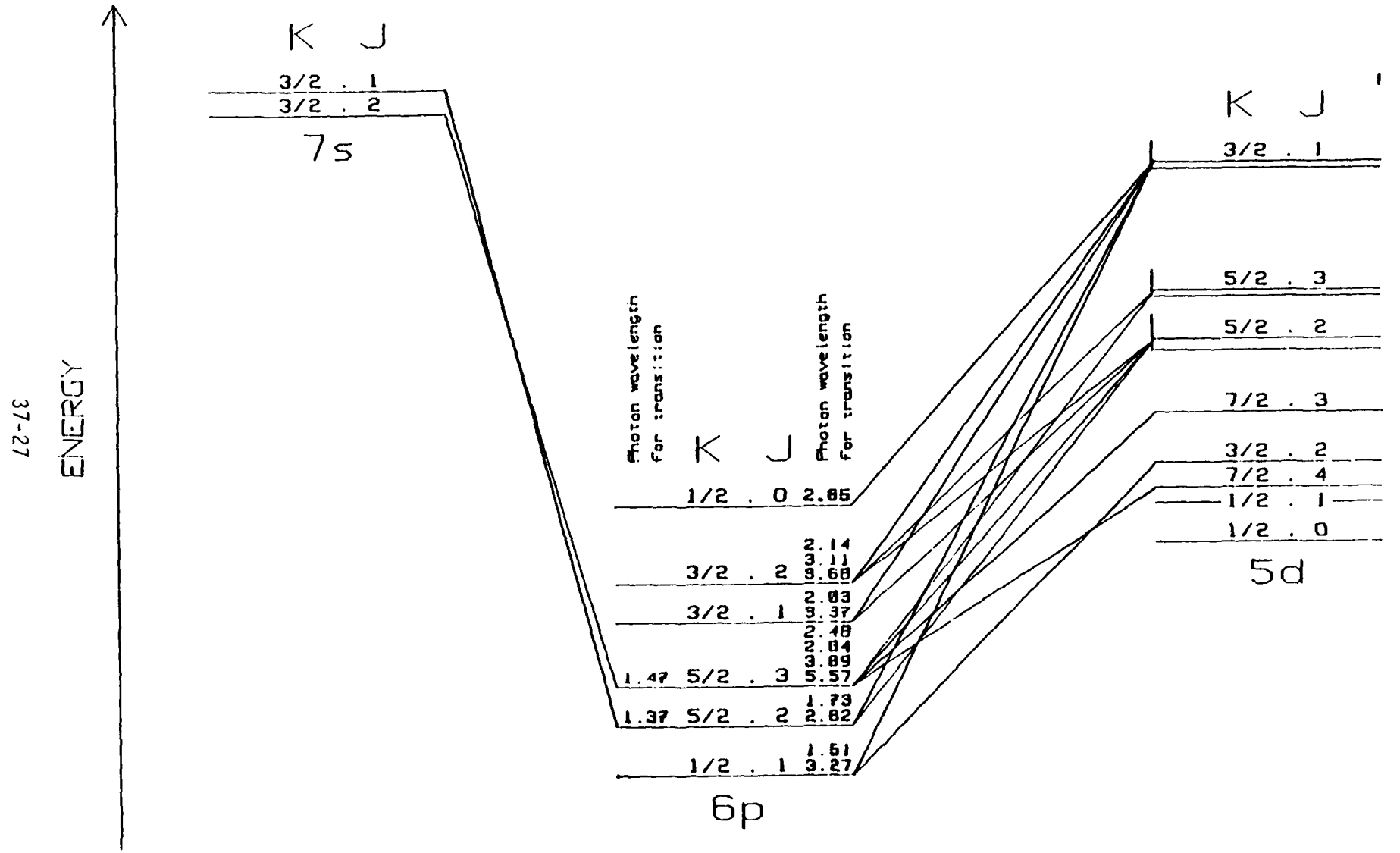


FIGURE 15. Partial Energy Level Diagram of Xenon.

Also, we have found that transitions originating from the 6p states are pressure dependent in an extremely anomolous manner (Ref 4). It would be of interest to examine these 5d states which strongly feed the 6p levels to see if they possess such abnormal pressure dependence.

Figures 16a through 16h are exemplary of data obtained. The largest wavelength observed was 3110 nm but it lay at the extreme of the useful range of the 2000 nm blazed grating and was quite weak. The largest pressure possible was 4 mTorr. Xenon possesses an extremely large ionization cross section. And as mentioned before, this results in an easily unstable beam condition in which a glow discharge can be created. Hence, one must choose the operating conditions of the electron beam carefully so as to maintain the integrity of the electron beam. We have included in the accompanying figures excitation functions at 4 mTorr (maximum pressure used) and 1 mTorr (minimum pressure used). Attention is given to the fact that the 7s(3/2,2) to 6p(5/2,3) transition is the only one we observed with significant pressure dependence (See figure 17).

Again, we have substantiated that with the present system, optical excitation cross sections 10^{-18} cm^2 or higher are obtainable out to 2650 nm.

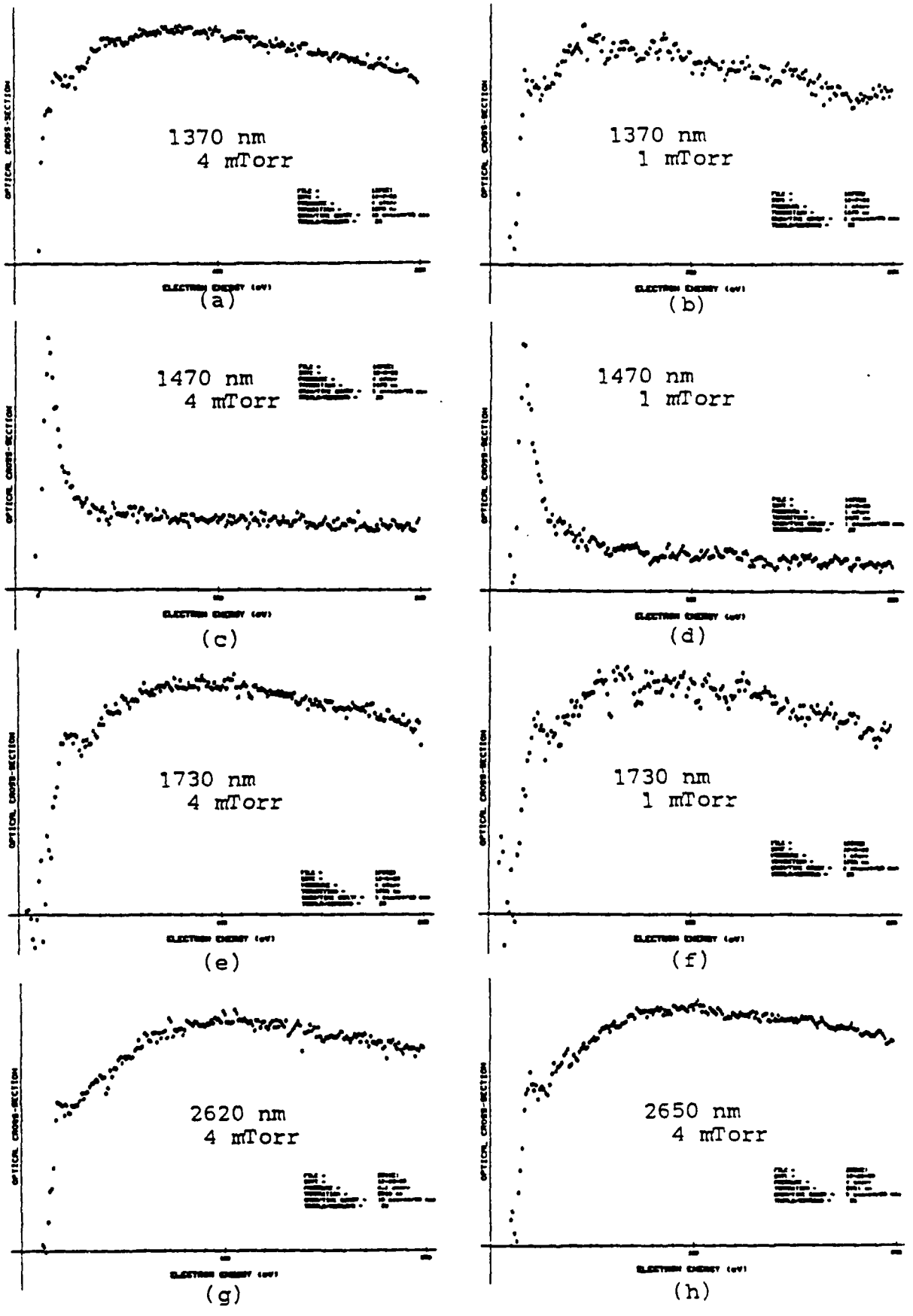


FIGURE 16. Xenon Excitation Functions(Optical) to 200 eV.

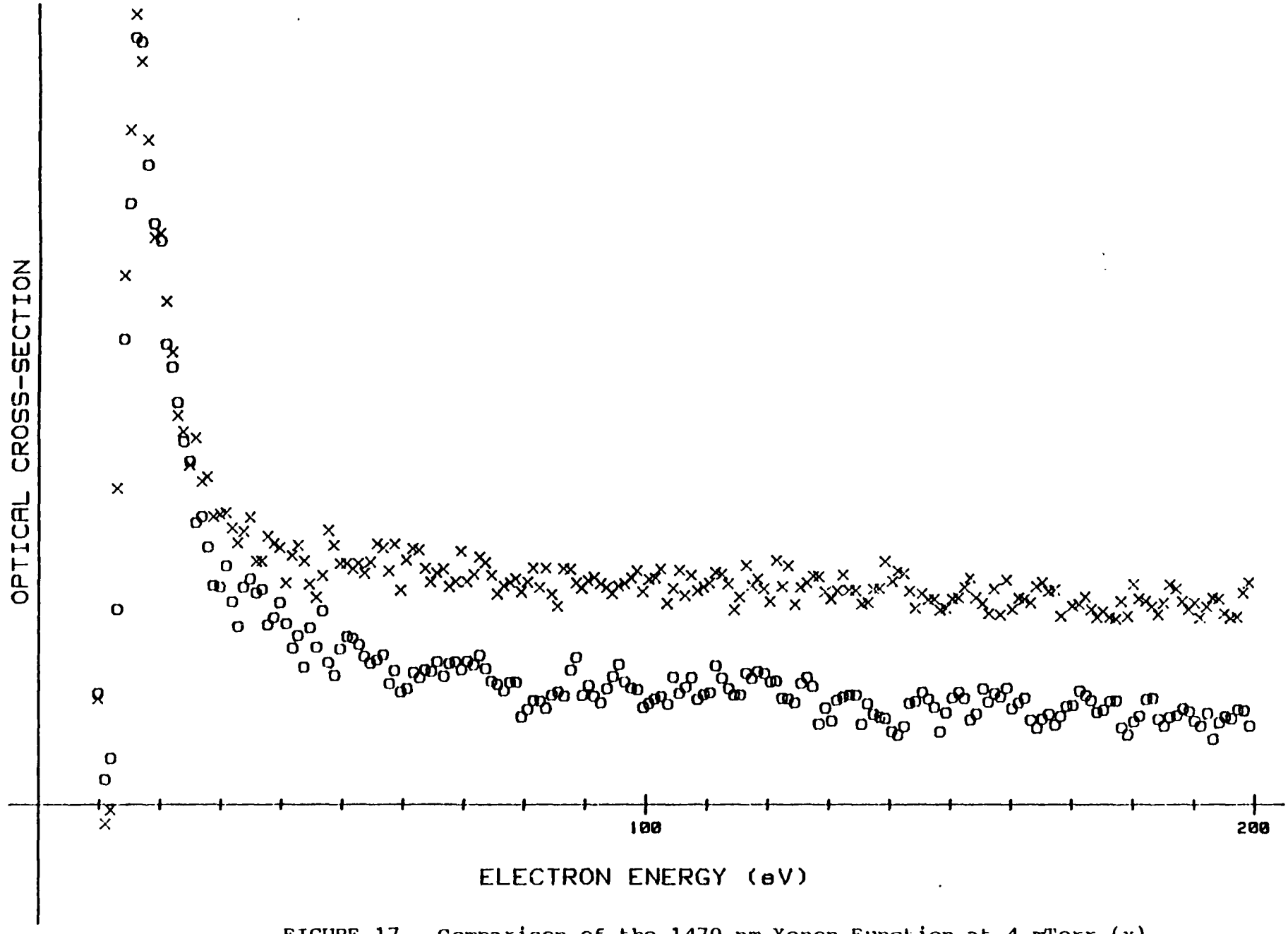


FIGURE 17. Comparison of the 1470 nm Xenon Function at 4 mTorr (x) and at 1 mTorr (o)

NEXT PHASE

As mentioned previously, at least 1 picoWatt must reach the detector when attached to the monochromator to obtain a SNR of 2. It is felt that we have closely optimized the detector and monochromator configurement. However, there are two improvements in the remainder of the system which should easily increase our detection capabilities by an order of magnitude.

1) Increase the solid angle from which radiation is gathered from the observation region. Presently, the geometry of the chamber is dictated by the electron gun design and thereby limits us to an f/12 speed. The monochromator and detecting element are capable of an f/3.8. We have designed and are presently building a new electron gun which will allow us to house it in a chamber that will enable us to 'get closer' to the electron beam. This new design will provide an f/4 value to almost match the detection system's optics. This should improve our signal by a factor of three (3).

2) Increase the electron beam from a 3 mm diameter beam to a 10 mm beam. Such an increase in beam size will result in an increase in beam current by a factor of ten (10). We decided to take this original data with a 3 mm beam since concern existed for the baffling of scattered light. The small cathode creates an amount

of background light that is significantly less than the 10 mm cathode dispenser. However, the baffling scheme we have designed and employed is so effective we feel confident that the increase in background will not be significantly greater than with the small cathode.

We have obtained a grant from Research Corporation to fund the above two improvements as well as to utilized calibrated spectral radiance sources to make absolute measurements of the excitation cross sections that are obtained. These modifications and measurements will be performed this coming summer, 1989.

REFERENCES

1. Tudor E. Jenkins, Optical Sensing Techniques and Signal Processing, Prentice/Hall International, UK.

2. H.W.S. Massey and E.H.S. Burhop, Electronic and Ionic Impact Phenomena, Oxford University Press, Oxford.

3. S. Lawton and T.A. DeTemple, Near Infrared Gas Lasers, Technical Report AFAPL-TR-78-107, WPAFB, Ohio (1978).

4. J.E. Gastineau, C.C. Lin, L.W. Anderson and K.G. Walker, "Electron Excitation Cross Sections of the 6p states of Xenon and their Pressure Dependence", Bull. of Am. Phys. Soc., University of Wisconsin, Madison, WI, April 1987, pg. 1156.

RESEARCH INITIATION PROGRAM

FINAL REPORT

SUBMITTED BY

BERYL L. BARBER

15 DECEMBER 1988

CONDUCTED BY

UNIVERSAL ENERGY SYSTEMS, INC.

FOR THE

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

ROME AIR DEVELOPMENT CENTER

GRIFFISS AIR FORCE BASE

ROME, NEW YORK

OTCP

SUPERCONDUCTOR TESTING

Recent developments on materials technology have been made that show super-conductivity at moderately low temperatures (liquid nitrogen). There is some promise that materials will be developed that will show super-conductivity at room temperature. Most measurements have been made using direct current and little is known of their AC characteristics. The main materials development has been that of small discs (approximately the size of a dime) and only direct current and antimagnetic phenomena have been studied. Manufacture has been difficult and the material samples are simple, such that it has been impractical to do RF (AC) characterization.

Even more recent developments in laser/implant techniques appear to be able to produce complex circuits of super-conductor material. These techniques should be able to produce microstrip and strip-line materials.

INTRODUCTION

This program will develop a set of techniques to both AC and DC characterize super-conductor material. Suitable fixture design will be provided. This will include at a minimum, fixture designs from .1 GHz to 18 GHz.

OBJECTIVE

Techniques will be developed to test and characterize super-conductor materials, using both DC and RF methods. Items investigated will include RF loss, skin effect, electromagnetic effects, power handling and contact resistance. Hardware designs will be made to provide for such measurements as circuit "Q" under RF conditions. The designs will also allow for simultaneous DC biasing to measure minimum and maximum operating levels.

The above designs will be made using strip-line, microstrip and coaxial methods. A circuit will be made, using cavity techniques to study "Q" and allow super-conductor samples to be tested in various "E" and "H" fields. Each test component will have provisions for cooling by either liquid helium or liquid nitrogen, or any other liquefied gas. Each will also operate at room temperature.

SUPERCONDUCTIVITY

Super conductivity has been with us many years. The most well known of early experiments was done by Meissner's group in Berlin in 1933. This magnetic-antimagnetic phenomena is the most widely demonstrated today. Recent developments in the technology have found superconductivity at ever increasing temperatures, some temperatures as high as 240 Kelvins (Wayne State University, Detroit, Michigan).

The designers of long power lines have dreamed of power networks using super-conducting carriers. The "Glitz"; however, is that even at liquid nitrogen temperatures (~77.1 Kelvins), the cost in energy of cooling the lines far exceeds the gain in efficiency. The mechanical stress problem is also a major factor. Most high temperature superconductors are very fragile.

Super conductor phenomena has been studied mostly at very low frequencies or direct current with the exception of the Josephson Junction (Brian Josephson 1962). The Josephson Junction has become a valuable tool in the laboratory, but as of this date has not found wide spread use in commercial (or military) products. I believe this is mainly due to the cost of cooling, not only in dollars but also in size and complexity.

CLASSIFICATION

Super-conductors have been roughly classed into two major categories. Class I super-conductors are those that exhibit a threshold and saturation effect. That is: Above some extremely small current they exhibit superconductivity. As the current increases it continues to operate as a super-conductor until at some level it "drops out" of super-conductivity.

Class II super-conductors exhibit superconductivity over a wide current range. Neither class has been accurately defined. There is a tendency today to class some material as super-conductors when they show a marked decrease in resistance even though it does not go to zero. It might be better to call these psuedo-super-conductors.

Most of our regular conductors, such as copper, silver and gold show a marked decrease in resistance when cooled to very low temperatures. The real advantage will occur when a strong, ductile material is found to be a true super-conductor (class II) at room temperature or above.

The theory of super-conductivity says that at some temperature the lines of magnetic flux internal in the material are repelled or forced to the outside. This is similar to the skin effect observed at high frequencies. In a class I superconductor this occurs until the flux density gets too great on the surface and super-conductivity ceases. This suggests that although R is equal to zero, the reactive component is not only not zero but will in fact increase.

SUPER CONDUCTIVE TESTING

Testing of super-conductivity has been done mostly at DC or very low frequencies. There is some work now being done in Japan and a few laboratories in the USA at very high Microwave Frequencies. Little is being done to characterize a super-conductor at any frequency. Qualitative observances are being made mainly at the low frequencies but there is a hesitation to make any quantitative definition.

Recent articles have been written on super-conductor resonant cavities. These cavities show results of Un-loaded Q's of 300. Varian Associates, in 1961 produced numerous uncooled cavities with Q's in the order of 40,000 and the VA-1280B had a minimum unloaded Q of 140,000.

In order to investigate the effect of super-conductivity it was decided to design a cavity with a unloaded Q of approximately 35,000. The VA-1280B design will be described later in order to compare designs.

The cavity design approach was picked at Ku-band at approximately 14 GHz. This allows for a small size and requires a smaller sample for test. The cavity is approximately 1.5 inches in diameter by 2 inches long. Both the input and output are on the base and the top is a temperature compensator. The base is loaded with lossey material to suppress spurious modes. The material of the cavity is invar and is silver and gold plated. Cavity operation is in the $Te_{1,1,1}$ mode. The cavity is designed in eight parts, a body base, top compensator, waveguide, two flanges and two tuning screws.

The VA-1280B is a very unique design. It operates in the $Te_{1,1,17}$ mode and has no spurious responses below 18 GHz. A copper wire (.062) is wrapped tightly around a cylindrical mandrel. A second mandrel is fitted snugly to the outside of the winding. Using vacuum techniques a mixture of carbon and epoxy is then "sucked" into the windings. After curing, the inside mandrel is removed and the outside mandrel and winding

are fitted into a lathe where the center of the winding is cut out untill there remains a surface with .010 inches between windings. See figure 1.

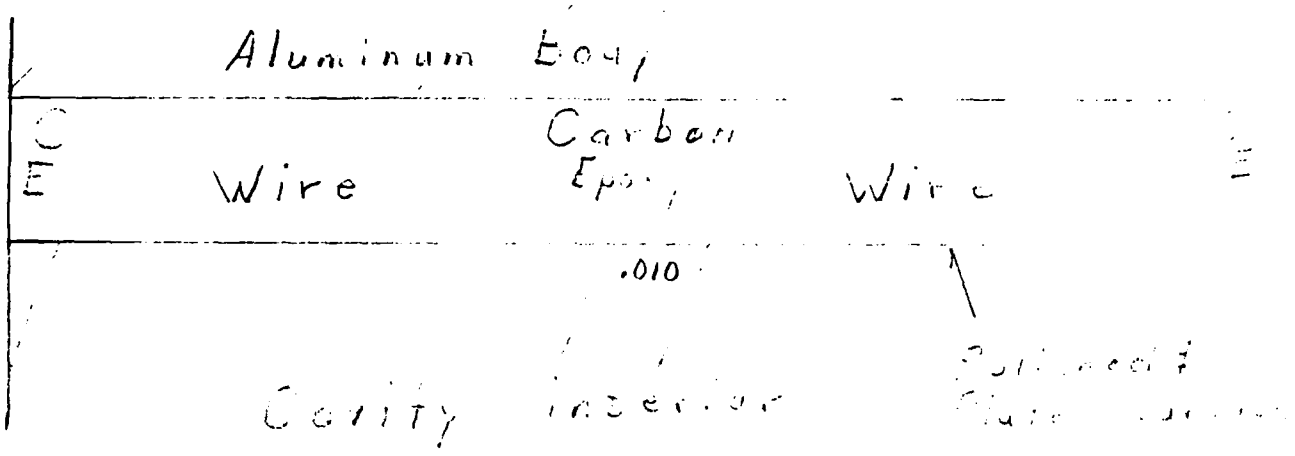


FIGURE 1

This gap is filled with lossy material and attenuates all spurious modes. The winding is then fitted into a cast supporting structure and plated with copper, silver and gold strike. The ends of the cavity are fitted with a tuner in the top. The lower end plate contains the coupling structure. If super-conductor technology gets very good, this type cavity would be an excellent test fixture.

CONTACT RESISTANCE

One of the most difficult problems using super-conductors is the input and output connections. As most connections are quite low loss, it is difficult to measure or compare this loss. Line impedance characteristics are easily measured but comparisons are difficult.

A long coaxial line was designed to make several measurements with the requirement that measurements be fast and reproduceable. Also desirable is the ability to D.C. or A.C. bias a super-conductor line and make RF measurements under bias conditions. Therefore "DC blocks" and Bias circuits were incorporated in the design.

To measure contact resistance, the line is 36 inches long. One hundred forty-four .25 inch super-conductor samples are fitted into a thin wall fused quartz tube. This becomes the center conductor of a coaxial line. Calibration of the line is done using a .062 brass tube, which is copper, silver and gold plated. The "outer" conductor of the line is a brass tube which is also copper, silver and gold plated using mandrel techniques to plate the interior.

When assembling either the calibration center conductor or the superconductor stack (or a single superconductor rod), "these"

are placed in the fused quartz tube. One end of the "Triaxial" fixture is assembled. The .032 diameter coiled wire is then fitted between the outer RF conductor and the gas manifold. The remaining end is then assembled and the outer insulating jacket installed. The device is now ready for test.

If biasing is desired, DC block/biasers may be used. These should be calibrated into the system before installing the fixture.

Each end uses a combination hair spring and titanium dioxide chip as a DC block/biaser. These are external to the main circuit and are calibrated into the test system. The circuit is as follows.

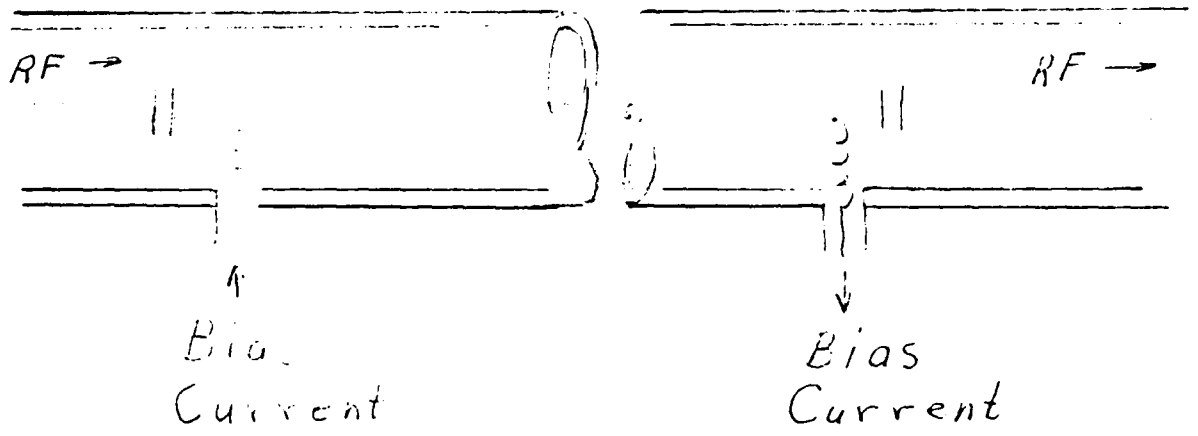


FIGURE 2

A larger brass tube is fixed in a triaxial configuration, with inlet and outlet connectors to provide cooling of the line. A throttling process assures that the internal gas remains at the prescribed temperature. Various gases such as liquid nitrogen, liquid helium, liquid oxygen and others may be used.

AMPLIFIER FIXTURE

Amplifiers, especially cooled amplifiers, require several connections to the external environment. These include input and output connectors, one or two bias connectors and input and output for the cooling media. A problem with the cooling media is that it must be kept at precisely the boiling point but not allowed to boil. In this amplifier fixture the input is designed using a throttling process. If the gas starts to boil, the throttling process reduces the temperature of the gas and reliquifies it. The fixture is also designed to automatically purge itself when a liquid media is applied.

When using the amplifier fixture numerous internal amplifier fixtures may be used. Dimensions have been made to accomodate a 1 inch by 1.5 inch thick film four to six stage amplifier. Smaller units, such as MMIC's may be tested by circuit board mounting.

SOLDERING TECHNIQUES

Line Connectors

Before plating the manifold parts (it is not absolutely necessary to plate but will maintain calibration longer if all parts are plated - remember we are measuring the difference in approximately one ohm and one hundredth of an ohm) the end female connector parts should be soldered on the manifold. Carefully place the connector part over the manifold making sure that the end of the manifold lines up with the inside base of the thread. Silver solder this in place on each end. The gas inlet and outlet tubes should also be silver soldered in their respective holes.

Ku-Band Cavity

The base, adaptor and two flanges should be silver soldered before final polish and plating. This may be done with a hydrogen torch on in a hydrogen oven. The oven technique is preferred but a torch technique is satisfactory but more difficult. Two 4-40 holes should be drilled and tapped after soldering. The part(s) is then polished and plated.

The compensator is then silver soldered to the cylinder, repolished and plated. This part(s) will not be soldered to the base section.

The gas inlet and exhaust pipes should also be silver soldered before plating. Note that the small pipes will be electrically at a pass frequency far above our operating frequency. i.e. beyond cut-off.

ASSEMBLY OF CAVITY

When assembling the cavity a specific technique should be used. Prepare the cavity by cleaning with acetone or alcohol. Then carefully spray the open end (inside) of the cylinder with teflon spray: Use very little - too little is better than too much. Now mix a small amount of epoxy with graphite powder. Any commercial epoxy will be satisfactory as this will be lossy. Carefully put a small ring of this around the outer notch of the base and install the cylinder part. This may then be removed after the epoxy has cured.

Several different techniques may be used to install the superconductor materials. A very thin cylinder inside the cylinder is preferred. Slight retuning (frequency change) will occur but Q meaasurements are the required results.

With the cavity in the system the phasing screws are adjusted to best couple to the input and output irises. Different lengths of cylinder may be made. Plus or minus 0.1 inches in length may be made without moding problems. The cavity, without superconductor, should tune at approximately 14GHz. If the source has harmonic content it may be necessary to put a small lossy card in the waveguide. This would be a piece of Mica (.010) resistance card cut .315 by .200, installed vertically in the waveguide, in front of the tuning screw, approximately one fourth of the distance across the waveguide.

NOTES FOR SUPERCONDUCTOR TEST LINE

Designs for three different lines are given. These are: a 36 inch line with APC-7 connectors with which to make measurements from D.C. to 12GHz, a 12 inch line with APC-7 connectors to make measurements from 2GHz to 12GHz and a 10 inch line with SMA connectors to make measurements from 2GHz to 20GHz. Assembly is the same for each line.

To assemble the line for calibration first install the brass center conductor in the fused quartz tube and place this in the center of the outer conductor. Put the gas-flow director (#11 copper wire) over the outer conductor, install the gas manifold over this assembly and assemble the connector on one end. Now carefully set the modified connector in the opposite end make sure the center pin is inside the quartz tube. This may be verified with an ohmmeter. Now install the retaining ring finger tight. After the installation is complete a length of winter foam pipe insulation may be placed over the manifold to maintain cold temperatures if reduced temperatures are desired. If cooling liquid is to be used connect the input and output gas lines. The end fittings are inter-changeable. The line is now ready for test. Replacing the center conductor with a superconductor will allow superconductor comparison.

APC-7 Connector Modification

Disassemble standard APC-7 bulkhead connector mounting, and machine (lathe) the square base to .500 diameter. The center pin is then cut to extend .050 past the base of the flange and the dielectric is flush with the flange. This will protect the fused quartz from compression breakage.

A threaded ring is now placed over the APC-7 mounting and the APC-7 connector reassembled.

OTHER LOSSES

As frequency goes up in the microwave range, copper or conductor losses increase less than do dielectric and contact losses. In fact, dielectric losses, in even the best insulators such as fused quartz, saphire and silicon are several orders of magnitude greater than copper or silver above 30 GHz. As a result of this, new MMIC technology using super-conductors has much better promise of success than thick film or hardwire technology. Comparisons of these circuits can be made in the amplifier test fixture. It is interesting to note that before the GasFet MMIC, the best noise figures were greater than 4db. at X-band. TI is presently delivering MMIC's at X-band with noise figures of 1db.

NOTES:

SUPER-CONDUCTOR TESTS.

Using a line capable of external biasing has produced some rather unusual results.

Looking at a transmission the impedance is $Z_0 = \sqrt{\frac{L}{C}}$

where the resultant Z_0 is a real number with no reactive component. If we measure the impedance of a line terminated in a $Z_0 = Z_L$ which is also a pure resistance, the reflection coefficient is zero. If $Z_L \neq Z_0$ but is a pure resistance, then the reflections coefficient has a reactive component at all points, except those occurring $\frac{n\lambda}{4}$ distances from the Z_L.

It appears that if a controlled current is passed through the super-conductor coaxial line the R component remains at zero but a reactive component increases. This appears to be an inductive component which suggests that it might be possible to tune active circuits in super-conductors using very small currents. Remembering that in a super-conductor R = 0 the necessary circuit could be quite simple.

Further studies need to be made, using newer super-conductors to verify these preliminary findings and to explore tuning possibilities.

BIBLIOGRAPHY

Far-Infrared Conductivity if High-Tc Superconductor $YBa_2Cu_3O_7$
Bonn D.A., Greedan J.E. et al
Physical Review Letters May 1987

Superconductivity
Eric Brus
Microwaves and RF July 1987

Superconductors Speed Picosecond Signal Analyzer
Jack Brown
Microwaves and RF April 1987

Superconductivity seen above the Boiling pont of Nitrogen
Physics Today April 1987

Transport and Structural Properties of the $Ho_1Ba_2Cu_3O_{9-\delta}$
superconductor Lee Sung-Ik et al
Applied Physics Letters, American Institute of
Physics July 1987

The Road to Superconducting Materials
Hulm, Kunzler and Matthias
Physics Today January 1981

BIBLIOGRAPHY

High-temperature superconductivity: what's here, what's near and what's unclear. by Karen Hartley
Science News v132-Aug 15'87 p106(3) 40L0789

Superconductivity glimpsed near 300K.
(superconductivity at room temperature) by Dietrick E. Thomsen.
Science News v132-July 4'87 p4(1) 40E567

Predicting new solids and superconductors. by Marvin L. Cohin. iL Science v234-)ct 31'86 p549(5)

High-powered discussionson high-temperature superconductivity. by Karen Hartley. Science News v132-Dec 5'87 p359(1)

The discovery of a class of high-temperature superconductors. by K. Alex Muller and J.Georg Bednorz iL Science v237-Sept 4'87 p1133(7)

Not-so-superconductors. (includes article on the theory of superconductivity) iL Economist v303-June 13'87 p93(3)

Superconductor claim raised to 94K: a joint effort by researchers at the University of Alabama in Huntsville and the University of Houston yields the first superconductor to operate above liquid nitrogen temperature. by Arthur L. Robinson
Science v235-March 6'87 p1137(2)

Analog superconducting electronics. by Paul L. Richards
iL Physics Today v39-March'86 p54(8)

BIBLIOGRAPHY

Carr, W. J., JR. AC Loss & Microscopic Theory of Superconductors. 170p. 1983. $70.00 (ISBN 0-677-05700-8). Gordon & Breach Science Publishers, Incorporated.

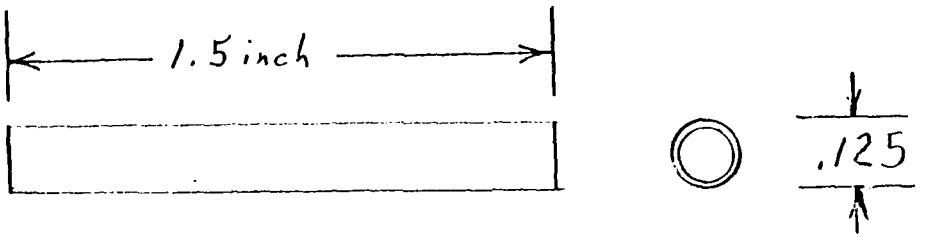
Geilikman, B. T. & Kresin, V. Z. Kinetic & Nonsteady-State Effects in Superconductors. 164p. 1974. Hardcover text edition. $36.00. (ISBN 0-7065-1428-9, Keter Pub Jerusalem). Coronet Books.

Horton, G. & Maradudin, A., editors. Dynamical Properties of Solids, Vol.3: Metals, Superconductors, Magnetic Materials & Liquids. 334p. 1980. $74.50. (ISBN 0-444-85314-6, North - Holland). Elsevier Science Publishing Company, Incorporated.

Huebener, R. P. Magnetic Flux Structures in Superconductors. (Springer Series in Solid State Sciences: Vol.6). (Illus.). 05/1979. $38.50. (ISBN 0-387-09213-7). Springer-Verlag New York, Incorporated.

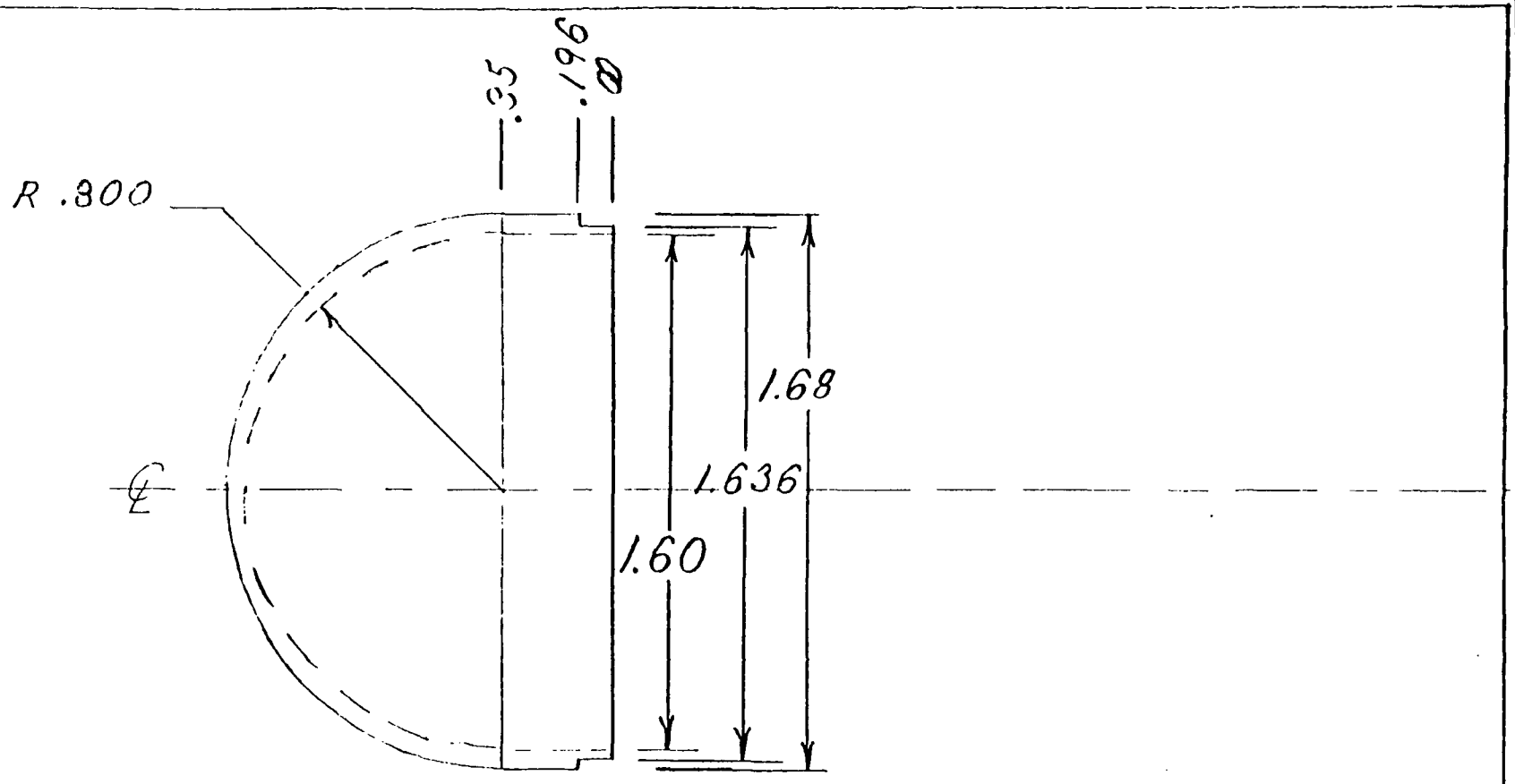
Moon, F. C., editor. Mechanics of Superconducting Structures. (AMD: No. 41). 137p. 1980. $24.00. (ISBN 0-686-69856-8, G00174). American Society of Mechanical Engineers.

Van Duzen, T. & Turner, O., editors. Principles of Superconductive Devices & Circuits. 370p. 04/1981. $41.50. (ISBN 0-444-0411-4). Elsevier Science Publishing Company, Incorporated.



GAS PIPE (CAVITY)
Material: Brass, 1/8 thin wall

C01



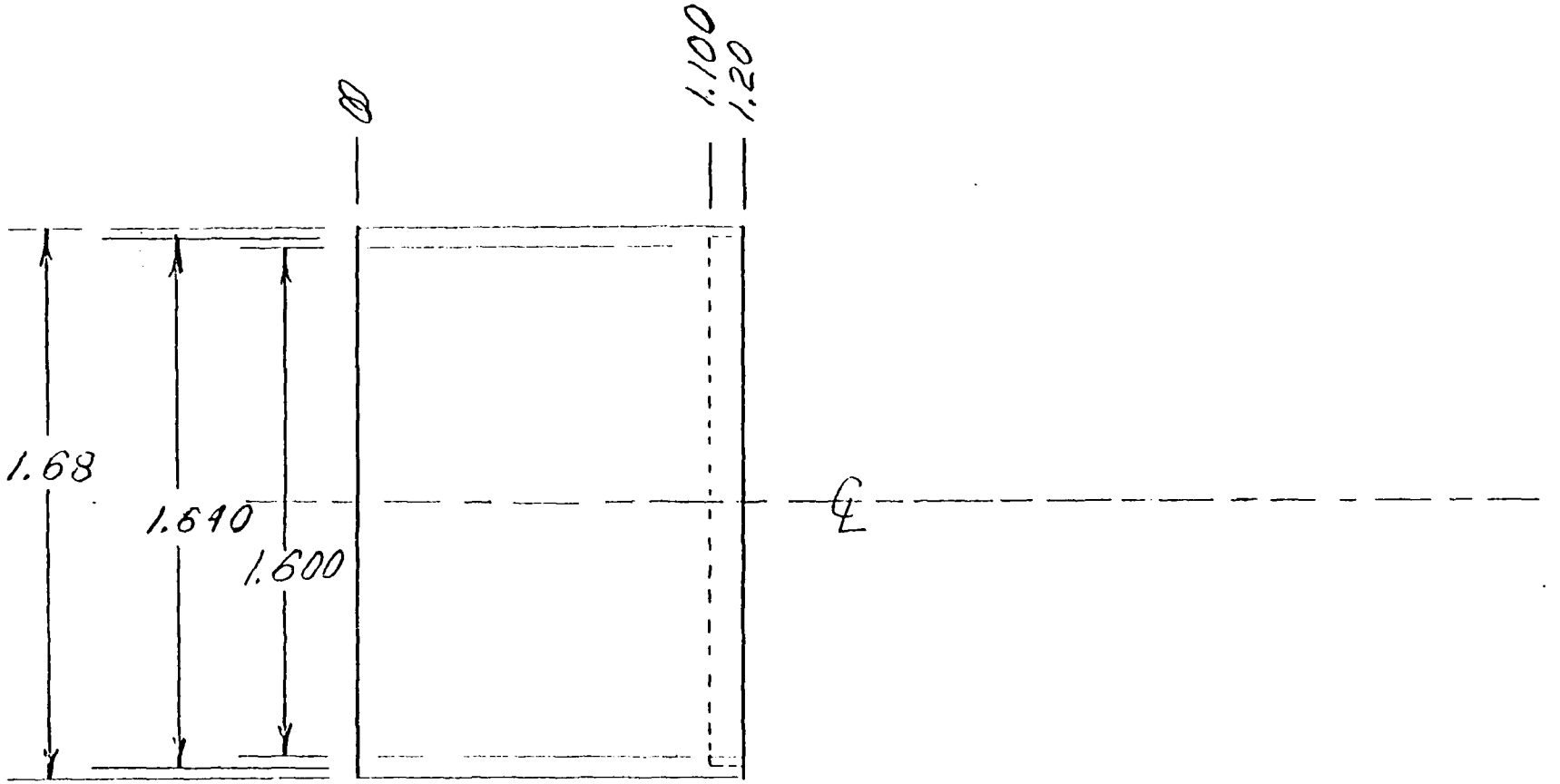
Notes:

1. Material: Invar
2. Inside Surface 2 μinch finish
   Make by pulling .060 Invar Sheet
   Polish Steel Ball and Case Harden
3. Plating: Copper Strike, .002 silver
   200 Micron Gold

1. .XX = ± .010
   .XXX = ± .002

CO2

KU-BAND CAVITY COMPENSATOR				



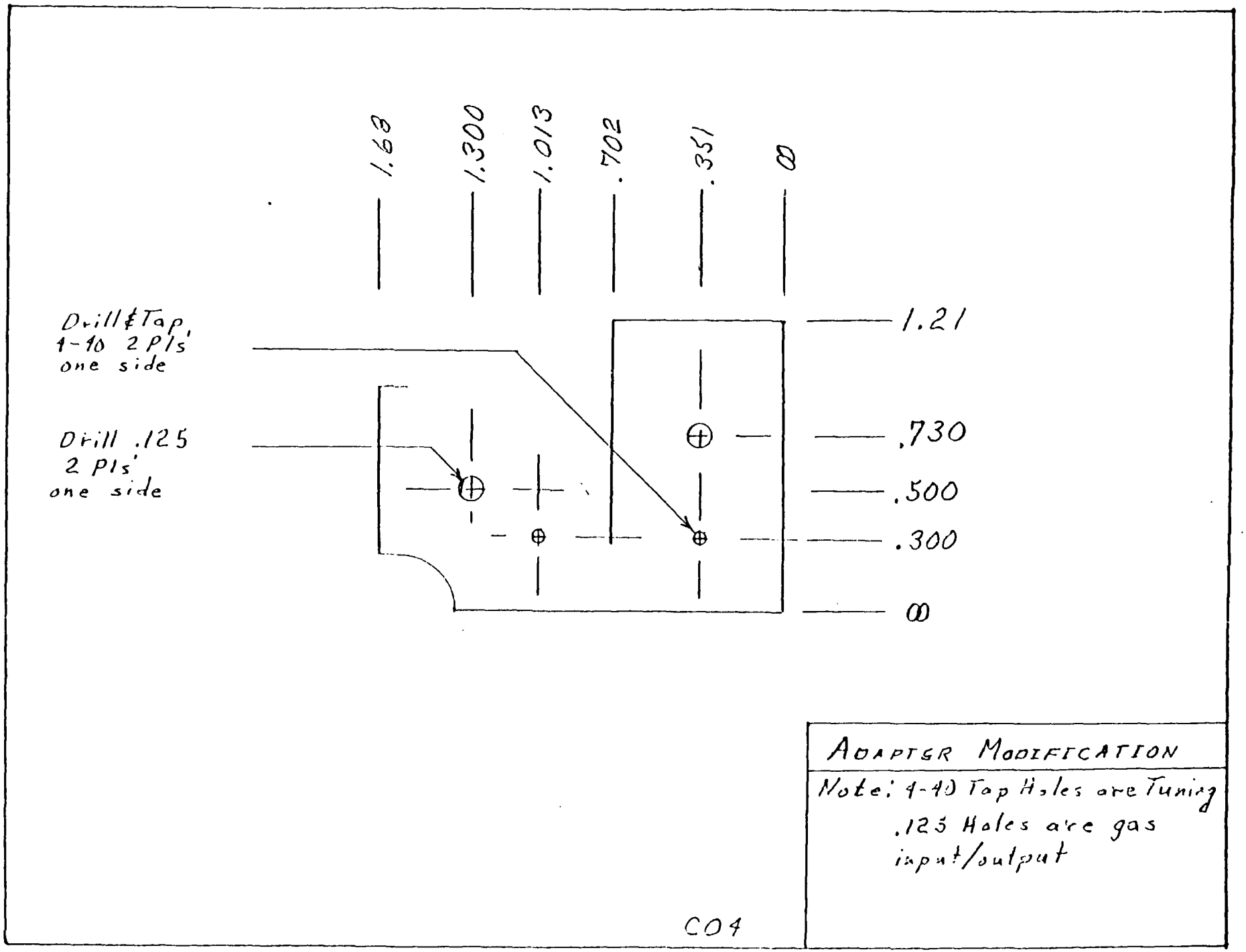
Notes:

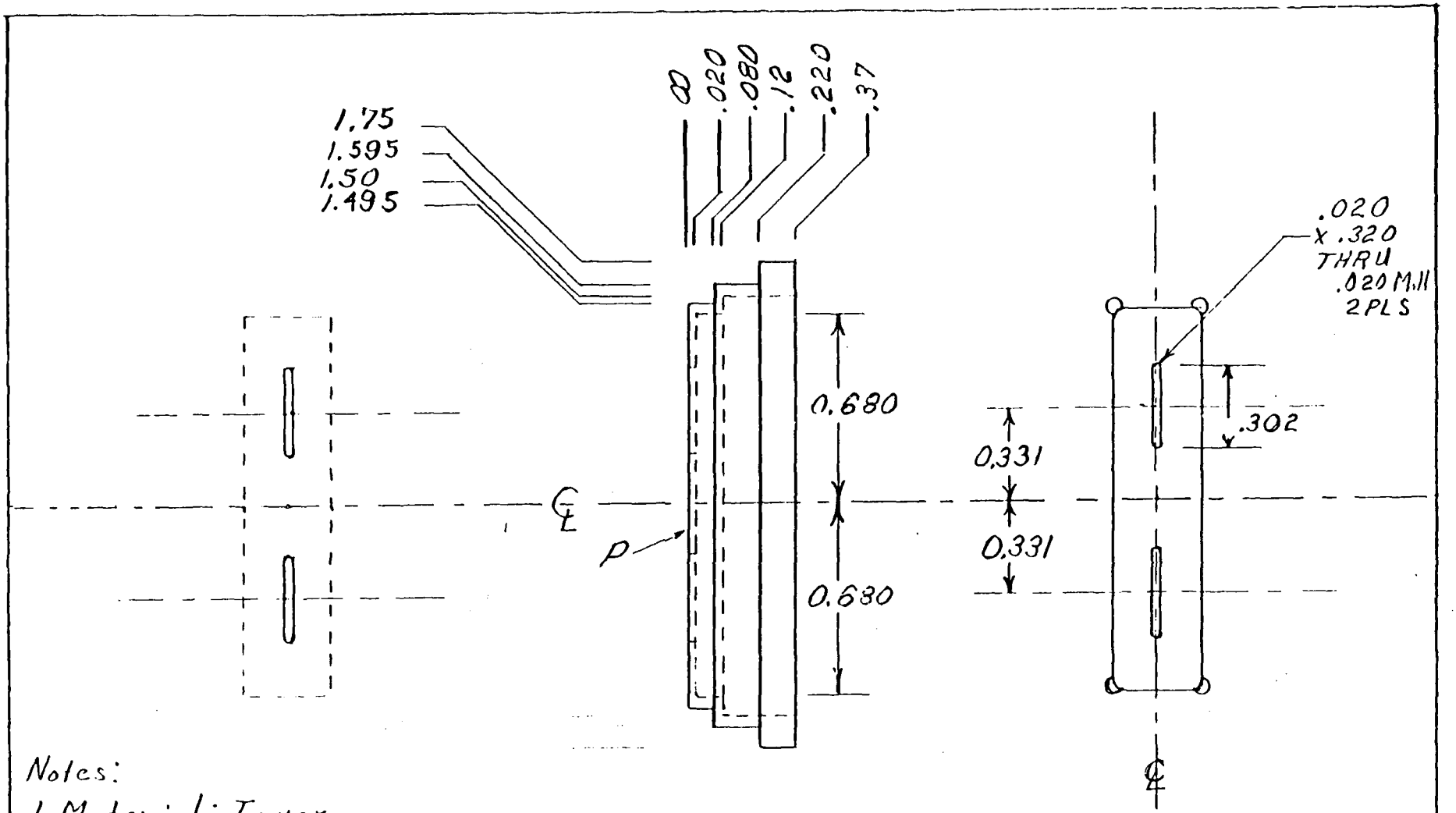
1. Material: Invar
2. Inside Surface 2 μinch finish
3. Plating: Copper Strike,
   .002 Silver
   200 micron Gold

1. .XX = ± .010
   .XXX = ± .002

C03

KU-BAND CAVITY CYLINDER				


1.68
1.300
1.013
.702
.351
00
Drill & Tap
4-40 2 Pls
one side
Drill .125
2 Pls
one side
1.21
.730
.500
.300
00
ADAPTER MODIFICATION
Note: 4-40 Tap Holes are Tuning
.125 Holes are gas
input/output
C04

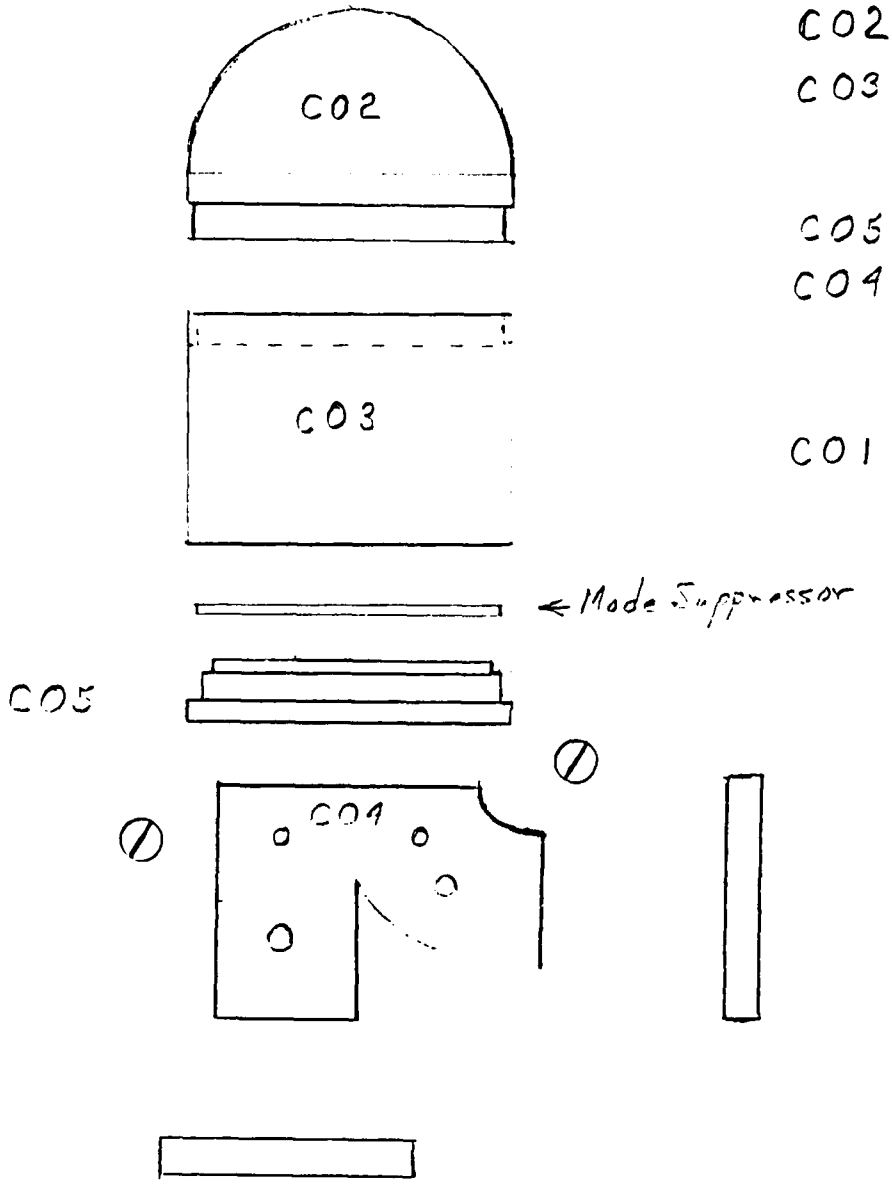


Notes:

1. Material: Invar
2. Surface P 2μinch finish
3. Other Surfaces 32μinch finish
1. Plating: Copper Strike,
   .002 Silver,
   200 micron Gold
5. .XX = ±.010
   .XXX = ±.001

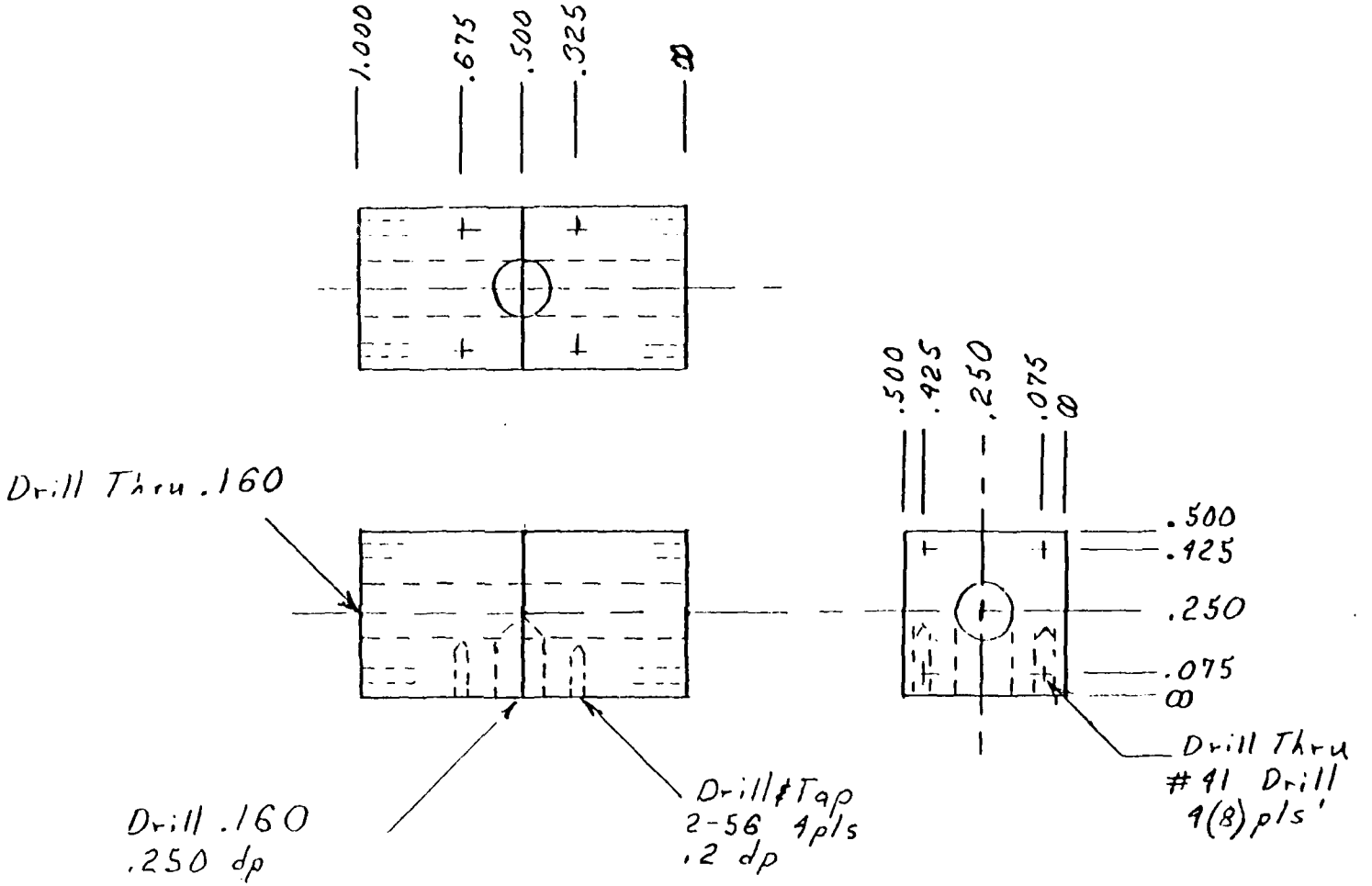
C05

KU-BAND CAVITY BASE



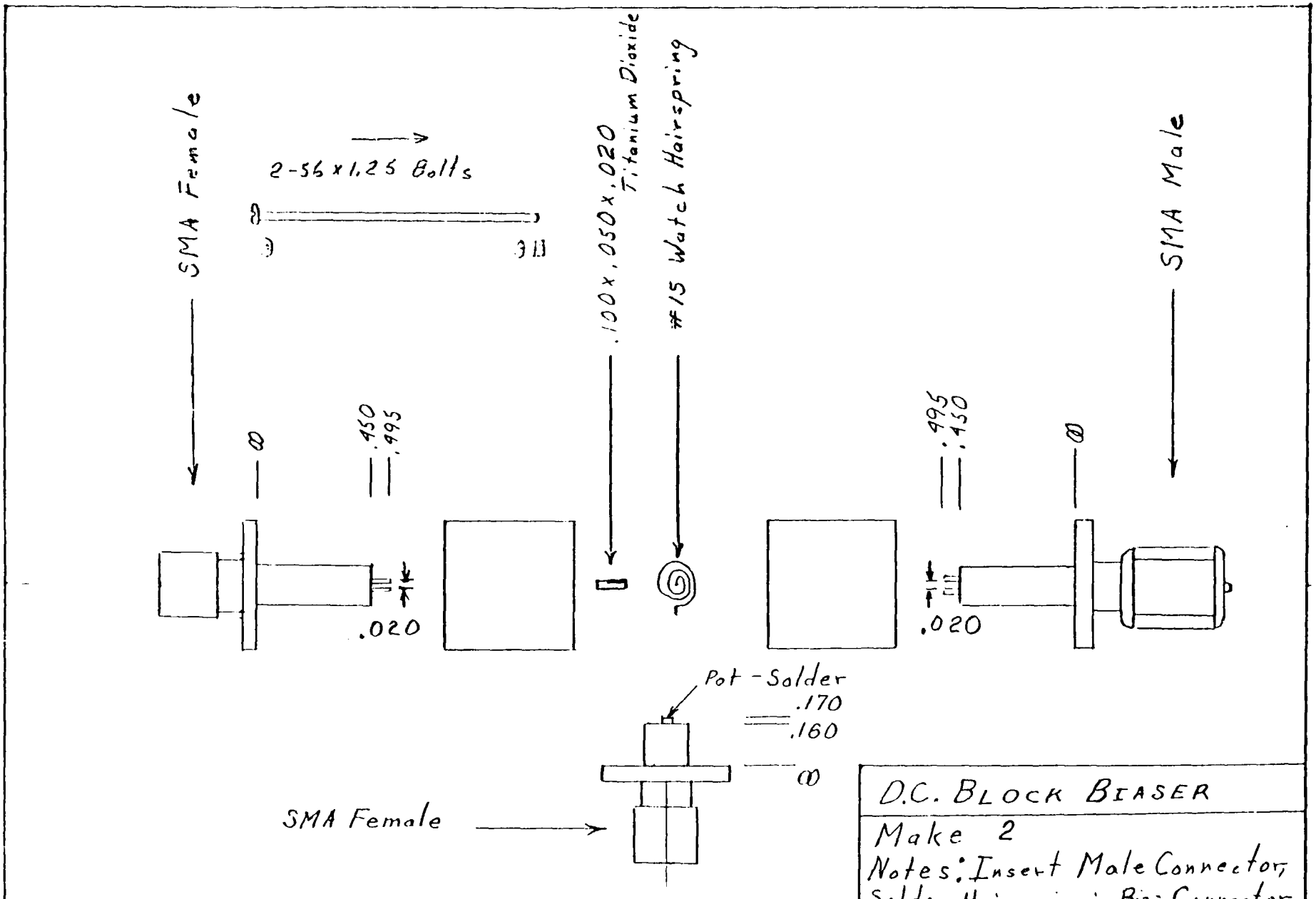
C02 Compensator
C03 Cylinder
Mode Suppressor
C05 Base
C04 Adapter
Phase Screw 4-40
Flange
C01 Gas Inlet Outlet

Cavity Assembly



Notes:
Material Aluminum 6061
Made two pieces .500 cubes
.xxx = ± .002
Plate if desired

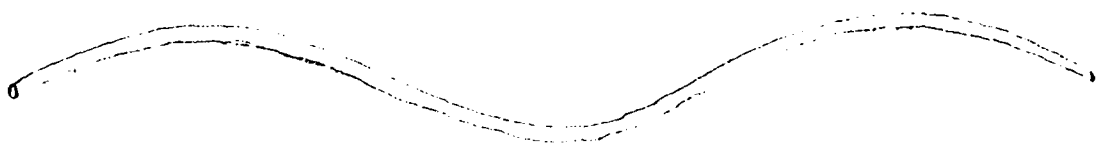
DC Block Biaser Body				



D.C. BLOCK BIASER

Make 2

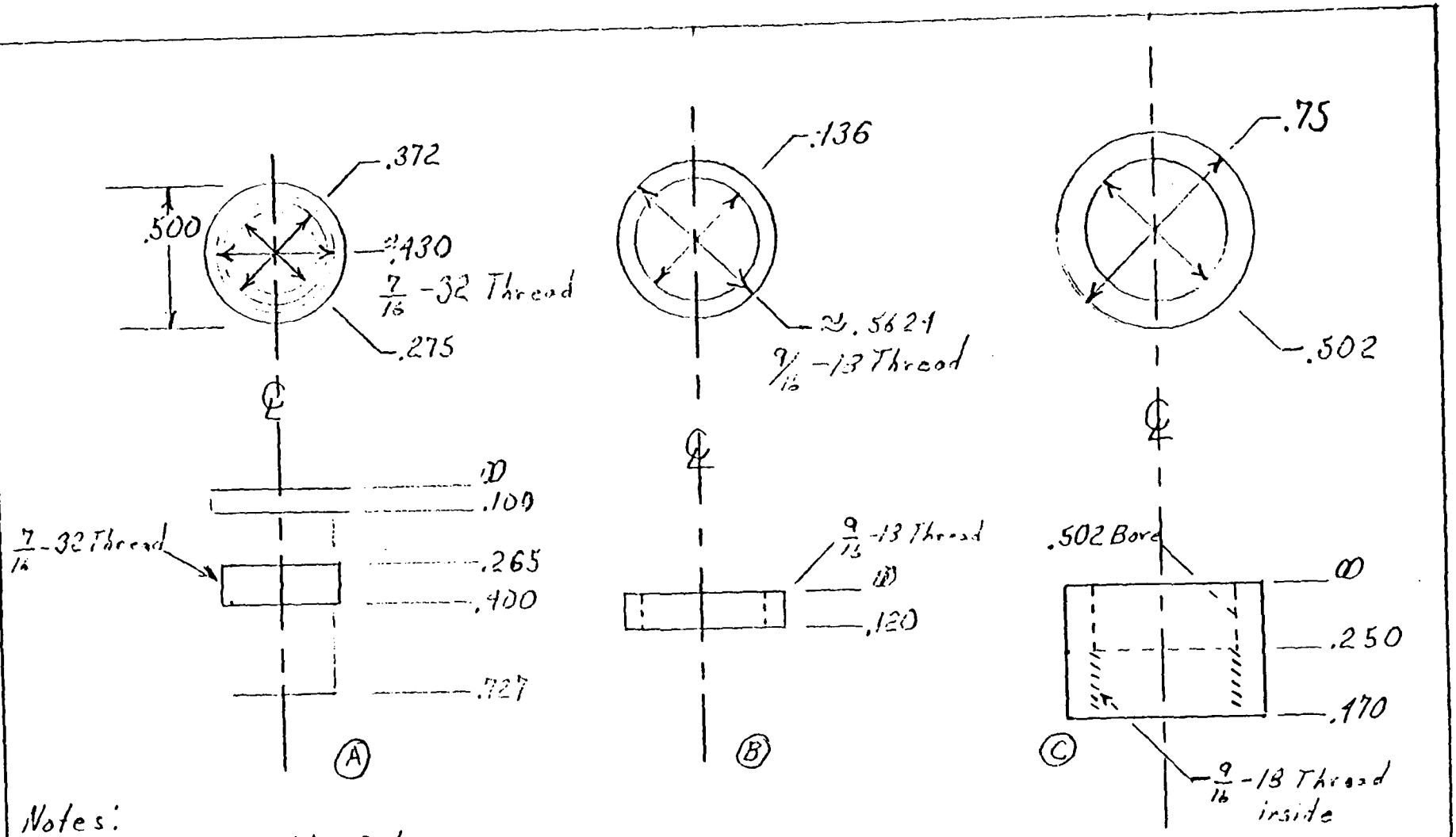
Notes: Insert Male Connector, Solder Hairspring in Bias Connector, Put Center Hairspring in male slot and hold with Dielectric. Install Female Connector and Bolts.



3-Turns

Gas-Flow Director				
Material	Size	Length	Length	Length
Copper	11-Gauge	42 inch	13 inch	15 inch

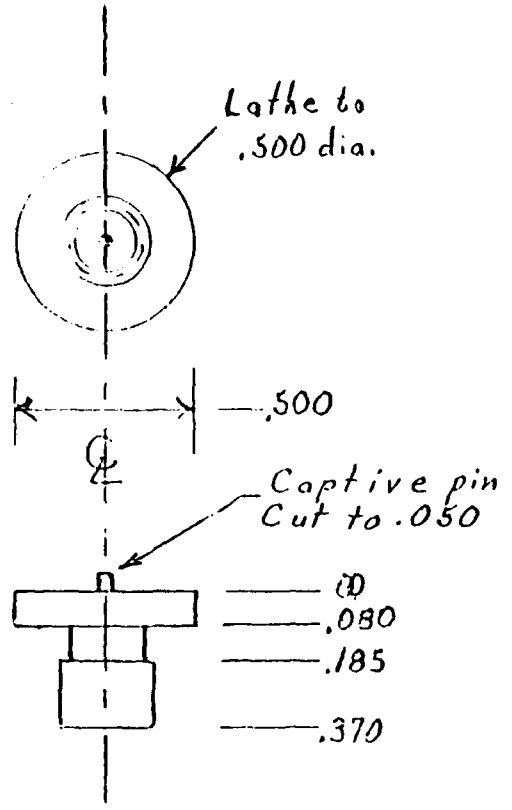
GAS PIPE
1/4" O.D. Thin Wall
1.0



Notes:

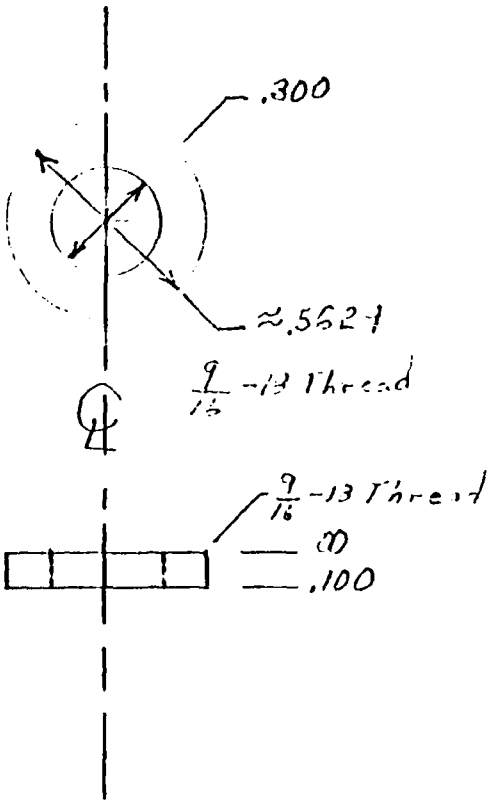
1. Material Nickle-Silver
2. A is modified Bulkhead Connector Body
3. .xx = ±.010
   .xxx = ±.002

APC-7 Connector Fitting				

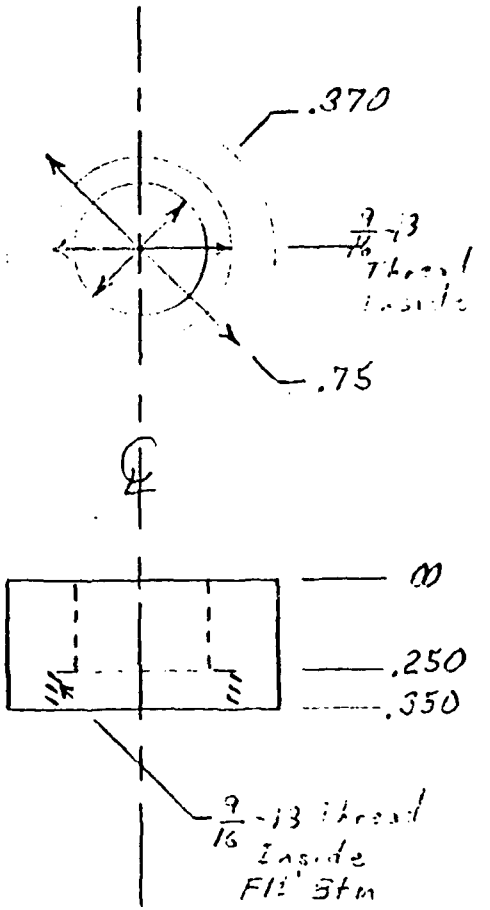


Modified SMA
1" Flange Connector

(A)



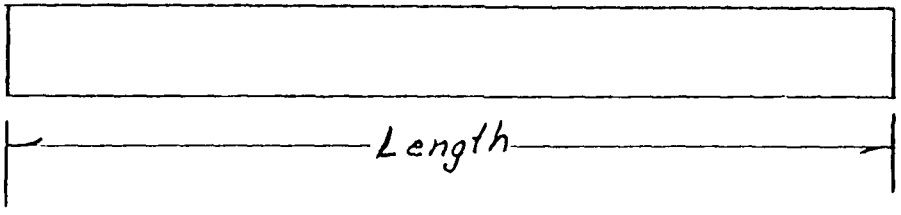
(B)



(C)

Notes:
1. .XX = ± .010
   .XXX = ± .002

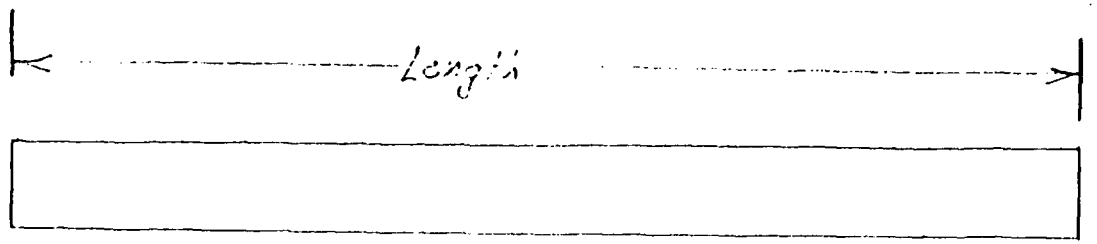
SMA Connector Fitting				
Part	Material	Pt. Num.		
A	SSteel	2755-6264		
B	Ni Silver			
C	Ni Silver			

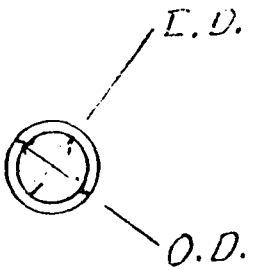


Dim. inches

.X - ±.05
.XX - ±.01
.XXX - ±.002

50Ω TEST LINE - APC-7				
Item	Material	Length	O.D.	I.D.
A	Brass	35.90	0.125	—
B	Brass	36.00	0.3	0.286
C	F. Quartz	36.00	0.175	0.130

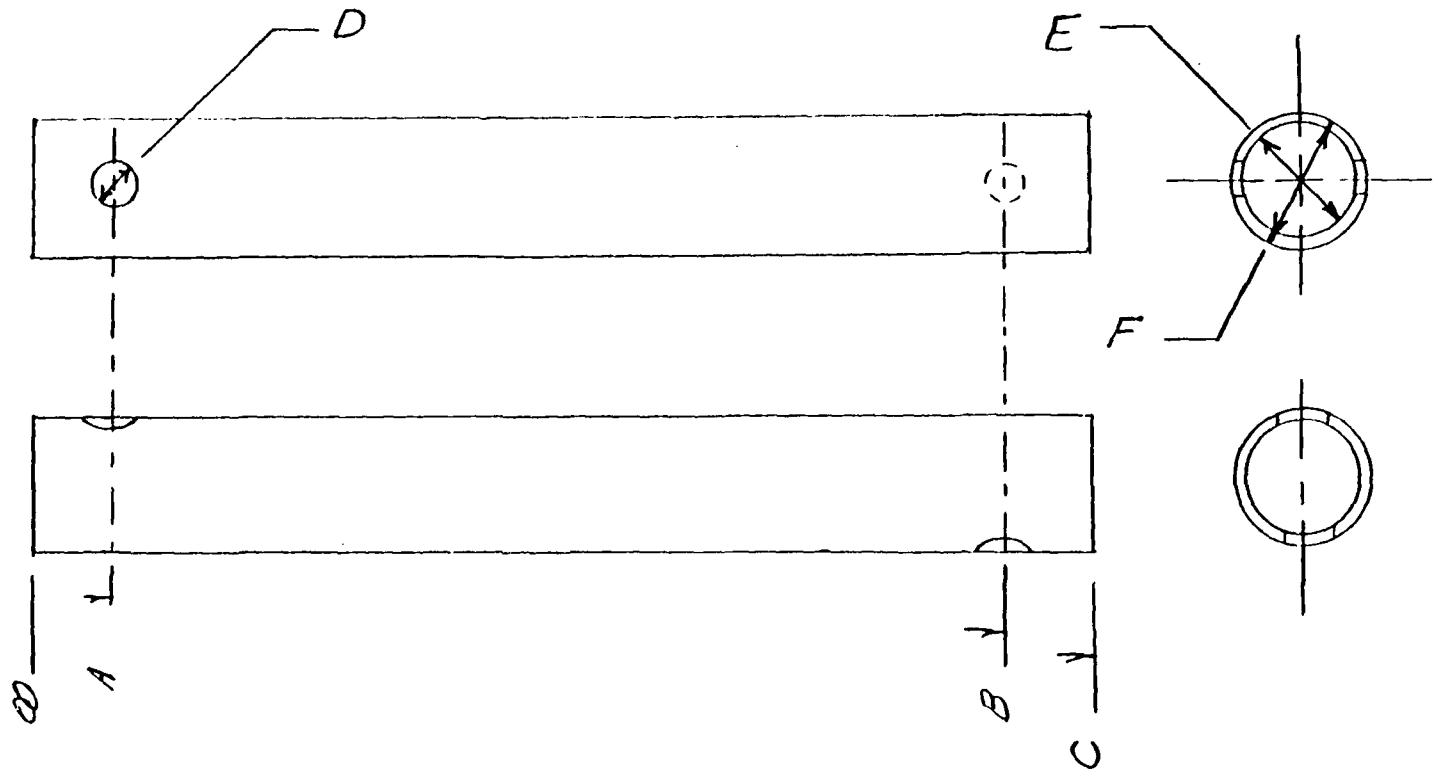




Dim. inches

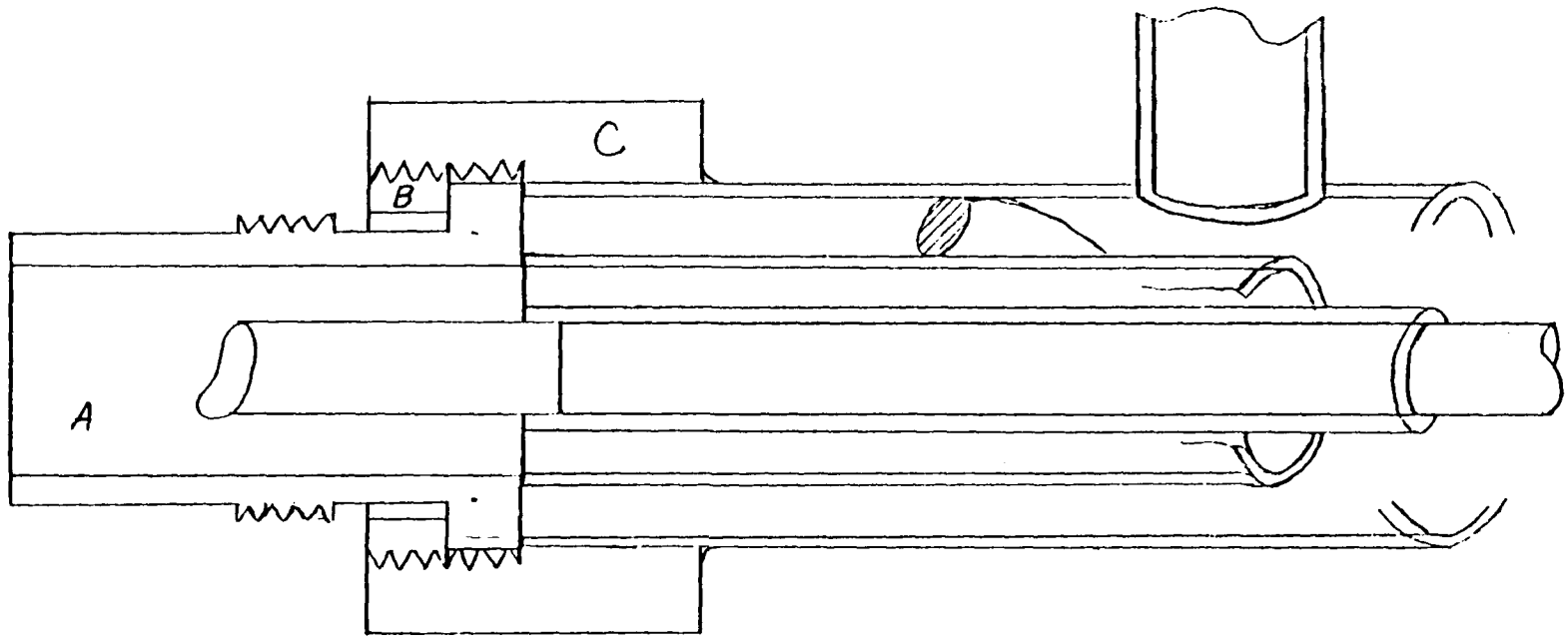
.x ±.05
.xx ±.01
.xxx ±.002

50Ω TEST LINE				
Item	Material	Length	O.D.	I.D.
Cen. Cond.	Brass	9.900	0.062	—
Out. Cond.	Brass	10.000	0.16	0.112
Pts. Hold.	F. Quartz	10.000	0.090	0.065

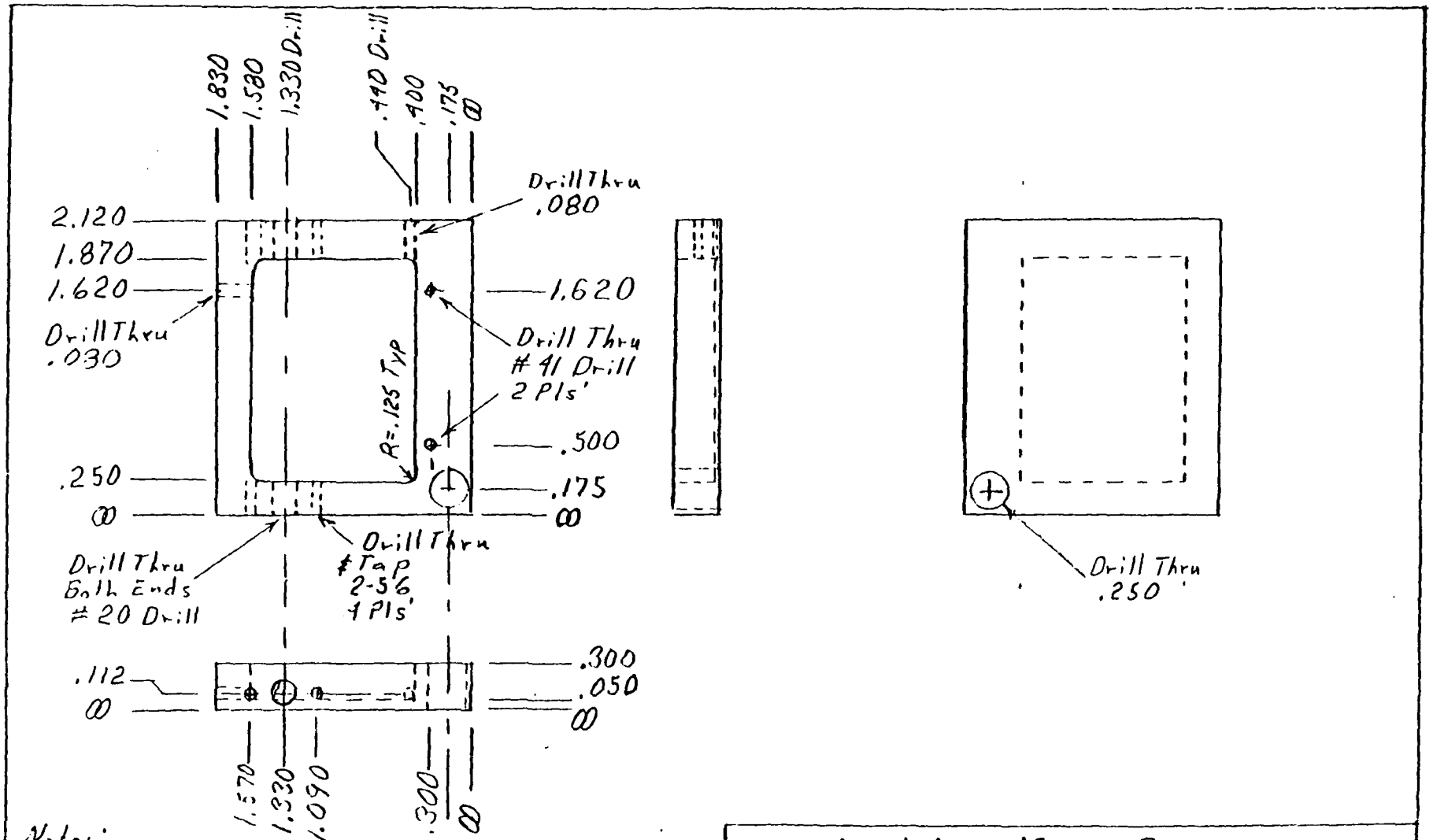


Mt'l Brass
Dim. inches
.X ± .05
.XX ± .01
.XXX ± .002

	GAS MANIFOLD					
	A	B	C	D	E	F
APC-7,	1.0	35.0	36.000	0.25	0.480	0.500
SMA	0.5	9.5	10.000	0.19–	0.34–	0.37–
APC-7$_2$	1.0	11.0	12.000	0.25	0.480	0.500

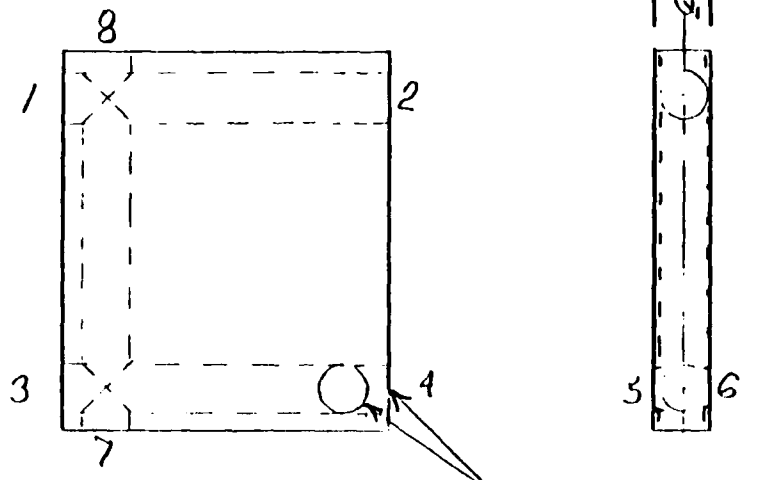


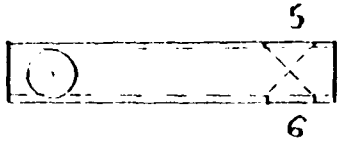
ASSEMBLY PICTORIAL

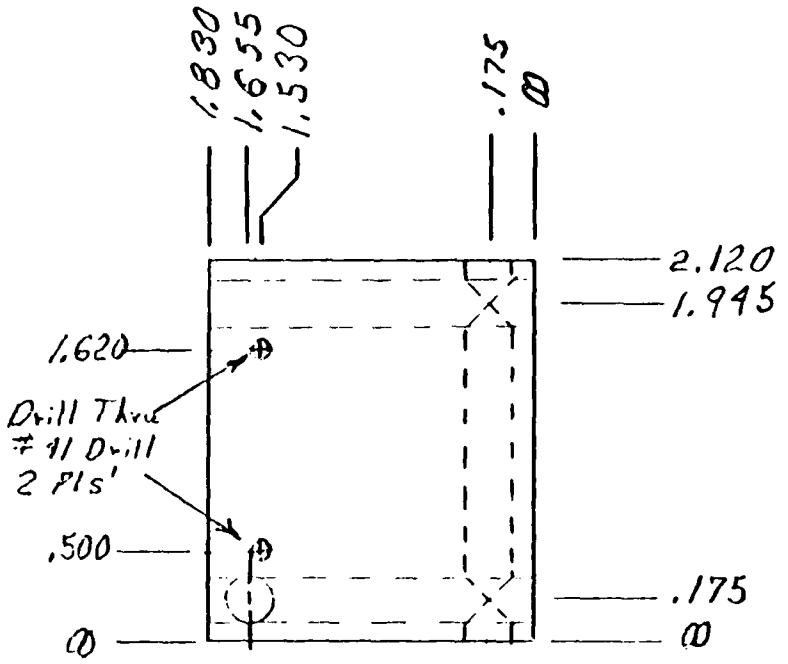


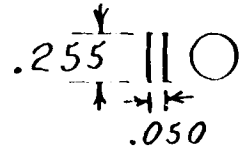
Notes:
1. Material: Aluminum 6061
2. Drill & Tap mtg holes 0-80
3. Plate: .002 Copper
   Gold Flash
4. .XX ±.01
   .XXX ±.002

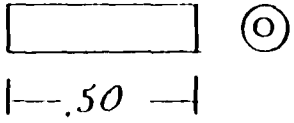
Amplifier Main Body				
SMA Con	App Spec	2950-6913		











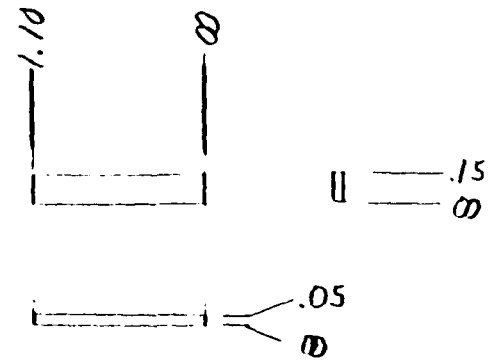
O.D. = .255
I.D. = .062

Notes:

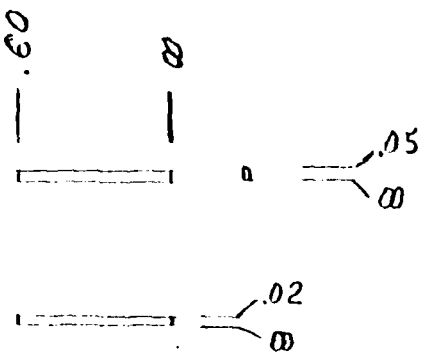
.XX = ± .01 inch
.XXX = ± .002 inch

Plate: Copper Strike
Silver .001 inch
Nickle Flash

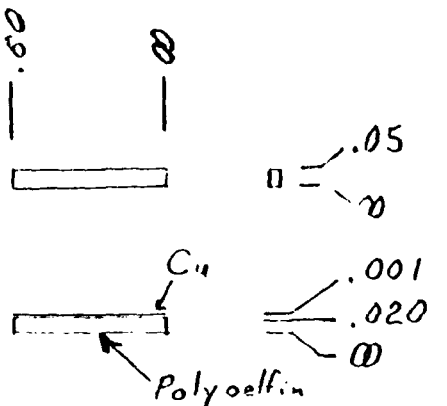
Amplifier Cooling Manifold				
Item	Plug	Pipe		
Top	123678	2		
Bottom	123578	2		



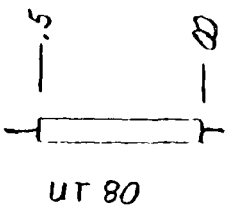
A



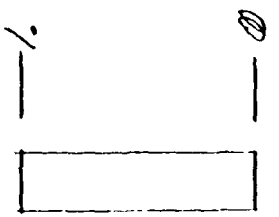
B



C



D

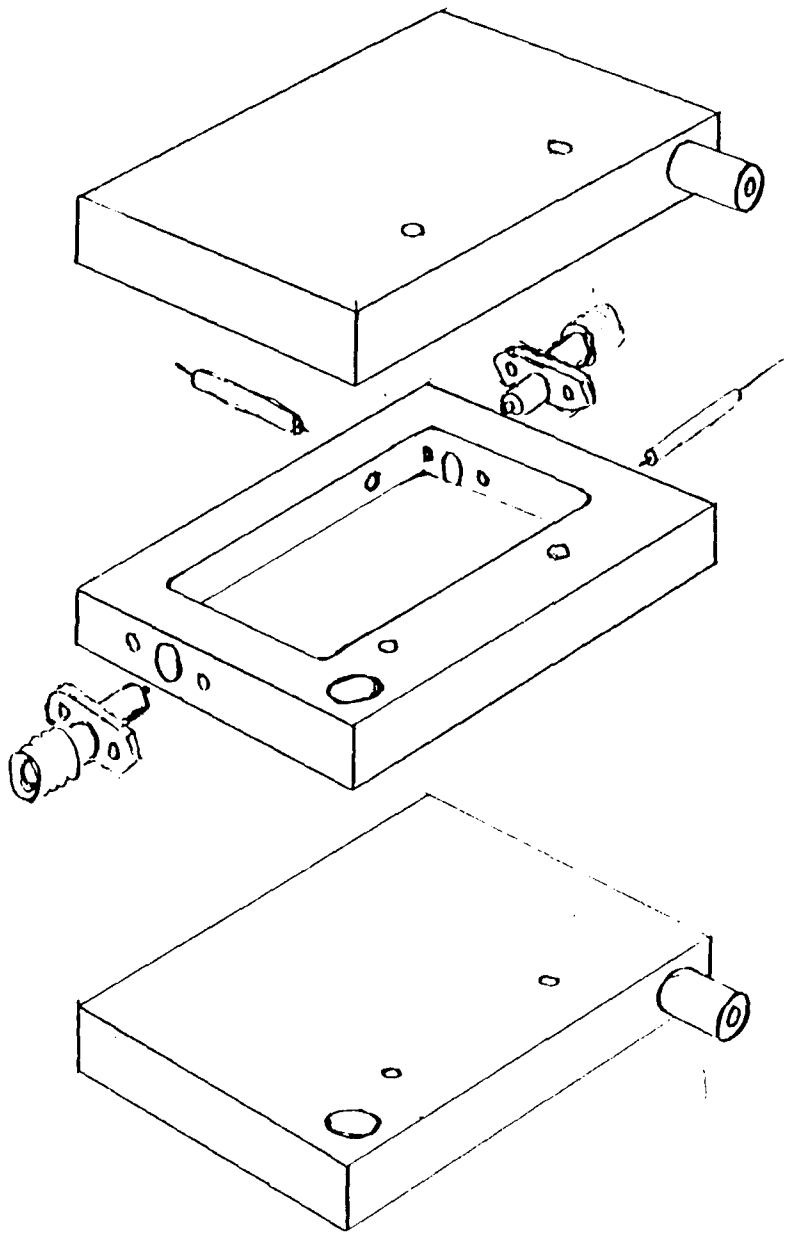


E

O.D. = .250

I.D. = .125

Small Parts				
Mode Sup	SiC		2ea	A
Insulator	TFE		2ea	B
Cool Pipe	Cu		2ea	E
Bias	UT 80		2ea	D



ASSEMBLY SKETCH

Cooled Superconductor Amplifier

Use 2ea 2-56 x1.00 Bolts

Final Report S-760-7MG-003:

A *Form and Function* Knowledge Representation for Reasoning about Classes and Instances of Objects

Professor Kevin Bowyer
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
(813) 974-3032
kwb@usf.edu

The period covered by this contract (S-760-7MG-003) approximately corresponds to the first stage of the three-stage project outlined in the proposal. The first stage of the project is primarily concerned with the development of a scheme for the representation of the geometric form of objects whose components may have non-rigid connections. Such a representation scheme has been developed and we are in the process of implementing an object editor module based on this representation scheme. The details of this representation scheme appear in a paper entitled "Representation of 3-D objects using non-rigid connection of components", presented at the SPIE 1988 Conference on Digital and Optical Shape Representation and Pattern Recognition. A short version of this paper was also presented at the 1988 Florida AI Research Symposium. A copy of each of these papers is attached to this report.

The other major effort during this contract period has been aimed at developing an appropriate means for blending information about *function* and *form* into a unified representation. This has proved to be substantially more difficult than developing a geometric representation which allows for non-rigid connection of components. We have worked out several possible unified representations, but have rejected them as being too cumbersome for use in reasoning about objects. The first alternative considered was to attach a procedural definition of a functional property of an object to a component of a generalized geometric model. This proved unworkable due to the great variability of geometric components which might provide the same functional property. The next alternative considered was to have separate procedural definitions of functional properties and geometric definitions of physical components and establish a graph structure relating the two. This proved unworkable due to the complexity of the processing needed to recognize an instance of an object. We are currently investigating a new alternative which would have functional properties as its primary (higher-level) elements and would expand a hierarchical structure into lower-level geometric descriptions of components as a particular object was interpreted. Once we have a detailed description of a workable unified representation for both form and function, we plan to apply (perhaps to AFOSR) for support to continue the project.

Proceedings of SPIE—The International Society for Optical Engineering

Volume 938

Digital and Optical Shape Representation and Pattern Recognition

Richard D. Juday
Chair/Editor

Sponsored by
SPIE—The International Society for Optical Engineering

Cooperating Organizations
Applied Optics Laboratory/New Mexico State University
Center for Applied Optics Studies/Rose-Hulman Institute of Technology
Center for Applied Optics/University of Alabama in Huntsville
Center for Electro-Optics/University of Dayton
Center for Optical Data Processing/Carnegie Mellon University
Center for Research in Electro-Optics and Lasers/University of Central Florida
Fiber & Electro-Optics Research Center/Virginia Polytechnic Institute and State University
Georgia Institute of Technology
Institute of Optics/University of Rochester
Optical Sciences Center/University of Arizona

4–6 April 1988
Orlando, Florida

Published by
SPIE—The International Society for Optical Engineering
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone 206/676-3290 (Pacific Time) • Telex 46-7053

Representation of 3-D objects using non-rigid connection of components

Louise Stark and Kevin W. Bowyer

Department of Computer Science and Engineering
University of South Florida
Tampa, FL 33620

ABSTRACT

Few three-dimensional object representation systems allow non-rigid connections between components of the object. We define a representation scheme that permits parameterized non-rigid connections, allowing one definition to specify a range of permissible configurations of an aggregate object. This representation can be used to generate 3-D instantiations of particular configurations, and 2-D projected images of particular 3-D instantiations. General issues involved in constructing such an object representation are outlined. The syntax of component description and connection types for our specific system is reviewed, along with the semantics of the allowable ranges of movement associated with each connection type. The actual representation system modules are also described.

1. INTRODUCTION

Descriptive representation of the physical form of 3-D solid objects is used in applications such as object recognition, CAD/CAM and graphics. Many different modeling techniques can be utilized to represent the object as a whole or to represent components (subparts) of the rigid object. Systems which allow objects to be defined as a connected structure of components must have a scheme to unambiguously specify the joining of components to form an aggregate object. Currently, most such systems allow only rigid connection between rigid components. We are investigating the use of non-rigid connections between rigid components. We plan to eventually explore representations for flexible components (e.g., cable) as well.

We envision an interesting and versatile representation of physical form to be one which allows:

> (1) definition of an "object" as a composition of multiple components, where
> (2) one component may be attached to another component by any of several types of "connection" (each connection type defines a different type of possible relative movement between the components).

By allowing parameterized non-rigid connections, the system can represent an object which can take on any of a range of allowable configurations. Major issues that must be considered with this type of representation include: 1) the syntax of component and connection type definitions; 2) the semantics of the allowable ranges of movement associated with given types of connection; and 3) how to interpret the aggregate object shape under different configurations.

System design trade-offs must be made concerning such issues as ease of implementation, ease of use, degree of realism in representing components and connections, and breadth of class of objects representable. Interrelationships between these issues cannot be ignored. The system described here models components as three-dimensional planar face solids. This representation is easy to implement and relatively easy to use. The trade-off is that it is not as realistic as might be. However, we feel this representation is adequate to model object prototypes for experiments on class representation.

Background information is presented in Section 2 in the form of related literature. System design issues and trade-offs are presented in Section 3. The breakdown of the actual representation system into individual working modules,

This work was supported by the AFOSR/UES Research Initiation Program under grant #S-760-00MG-003.

along with the description of connection types and how they are defined is covered in Section 4. Example objects (in this case, chairs) are depicted along with their directed graph structure and attributes. A summary and recommendations for further research concludes the paper in Section 5.

2. RELATED WORK

Current literature on solid modeling deals mainly with rigid solid objects. Good surveys of the literature in this area can be found in [Besl85] and [Requ80, Requ82]. Modeling of parameterized objects has been addressed by only a few researchers. The systems reviewed in this section have incorporated the idea of objects whose components may have a range of allowable orientations.

One notable representation scheme which allows flexible connections between components is described by Nevatia and Binford [Neva77]. Their representation allows objects to be defined as a combination of simpler subparts. The subparts are modeled as generalized cylinders, and are connected at "joints". The object is described by "connectivity relations, descriptions of the individual parts and joints, and global properties of the object" [Neva77]. Two or more subparts, stored as an ordered list, can connect at a joint. The angular and size relationships of the subparts listed decide the joint type. Some parts are allowed to be articulated, depending upon the type of joint defined. It is pointed out that, for the system described, "articulations of parts of an object are assumed to be completely unrestricted" and that without specified articulation limits, reliable discrimination between similar objects is difficult [Neva77].

Brooks and Binford describe a representation, used in ACRONYM, which allows modeling of specific instances, subclasses, and classes of objects [Bro81a, Bro81b]. The examples in [Bro81a] are oriented toward industrial parts, while the examples in [Bro81b] are oriented toward wide-bodied passenger jet aircraft. Their representation allows for variations in size and structure between specific instances within the class. Classes of objects are represented by a range of allowable variations. To model a specific rigid object they completely specify the constraints, essentially narrowing the range to a single value. Therefore, they do "not need to distinguish between specialization of the general model to a subclass, or an individual" [Bro81a]. While component connections may occur at different places for different object subclasses or instances, all connections are assumed to be rigid. Flexible connections between components are not discussed.

Brooks [Bro81c] further describes the internal representation of the volume elements of the geometric objects in ACRONYM, which uses a frame based system. The components of objects, represented as generalized cones, are entered by the user in a hierarchical order. An affixment tree of these components is generated that describes the object from coarse to fine levels of detail. Brooks notes that affixment does not denote attachment and that in some cases this could cause problems. The geometric models can be represented with variations in size, structure and spatial relations. Sets of constraints on these variations represent classes of objects; added constraints produce subclasses or instances.

Grimson deals with "families of objects that are characterized by sets of free parameters" [Grim87]. Families of parameterized objects can change by a scale factor, have rotating subparts, or subparts that stretch. Object models are represented by sets of planar faces. The set of faces for a model is not restricted to be connected or complete. One class of objects demonstrated (scissors) allows a limited number of moving parts with a single degree of freedom. The point of rotation is placed at the origin of the model coordinate system. This means that a model can have only one join of components, and that this join allows rotational movement. Stretching deformation was allowed for a family of hammers whose handles can stretch along the axis of the handle. The stretching axis is aligned with the x-axis in model coordinates. The idea of multiple types of connection for a single object is not addressed. Model edges are matched to sensory data edges, solving for transformation angle, scale factor and translation vector. All geometric constraints developed for the search process are for 2-D data.

Badler and O'Rourke modeled objects that are representations of quasi-rigid segments connected at articulable point-like joints [Badl79]. The human body is modeled using "a representation for the object segments and joints, the surface and coordinate system of each segment, and well-defined mathematical relations between adjacent segment coordinate systems during movement" [Badl79]. Overlapping spheres are used to model the "skin" of the segments. The skeletal frame of the human body lends itself to a tree structure where nodes depict the point-like joints. Each branch of the tree can be associated to the rigid bones known as segments. A mathematical relation is set up between adjacent

segment coordinate systems. A "standard position" is established relative to the ground. Possible limitations of twist are stored in a record definition for each segment. Default orientations, which are associated with the natural positioning of the limbs, are established through a standard orientation function for each limb. In this way, physical limitations of joints can guide a change in direction of twist during movement. An override of standard orientation requires further adjustment. An instantiation of the body is established "with respect to the environment through a chain of special instances of joints and segments" [Badl79].

Goldberg and Lowe [Gold87] base the representation used in SCERPO on an affixment tree structure introduced by Brooks in ACRONYM [Bro81c]. All parts of a model are related to a camera coordinate system through a "sequence of constant and variable rotations and translations" [Gold87]. Components are represented by edge descriptors. Goldberg states that the parameterized connections deal with translation, rotation about major axes, scaling and stretching. An example depicted shows a stapler that is allowed rotational movement of two components about the same axis relative to a third. A translational movement of one component relative to another is also allowed.

Ponce et al. [Ponc87] describe the geometric modeler of the Successor (to ACRONYM) vision system which uses a straight homogeneous generalized cylinder representation of components. Composite objects are formed using set operations. An object's volumetric structure is maintained in a binary tree structure called an assembly tree. Subobjects' relative positions are represented by an affixment graph whose arcs represent geometric transformations between primitives. Affixments are specified "by spatial relationships between planar faces of primitives" [Ponc87]. Affixment parameters can be assigned symbolic values (variables and s-expressions). An extension to the system is planned to allow generic modeling.

The use of parts or components to model objects has been used by many researchers to aid in the explanation of human perception. Recognition-by-components (RBC) is a theory proposed by Biederman in which objects are recognized by breaking them into a set of simple geometric components ("geons") [Bied87]. Thirty six qualitatively different generalized cylinder primitives comprise a set of geons which is proposed to be the set of primitive elements used to represent objects. Maintaining relations between geons in the representation of an object is very important. Two solutions of how to represent non-rigid objects are discussed. One solution suggests modeling or setting up structural descriptions for each arrangement of an object's geons of sizable difference. A second solution suggests specifying a range of possible relational values between components.

One of Pentland's representational goals is "to describe scene structure at a scale that is more like our naive perceptual notion of a 'part' " [Pent86]. Pentland discusses modeling natural forms and artifacts. The representation is based on superquadrics and fractals, which allows simple composition of components. The modeling primitives are associated to the "parts" of the form or artifact. Pentland believes that a more general vision system would follow the paradigm of a model base consisting of "parts that make up the specific object, rather than a model of the entire object, and the goal is to identify those component parts" [Pent86]. Primitives are joined by Boolean operators. By using a 3-D modeling system called "SuperSketch" users can create models of scenes of natural form. Pentland's work is directed toward understanding representation of natural form and human perception. Further research points include expanding the set of process-oriented modeling primitives to include such things as branching structures and particle systems.

3. CHOICE OF REPRESENTATION SCHEME

There are many important issues to be considered when designing a representation scheme, and trade-offs to be made with each design decision. Priorities concerning the type of information needed in the representation scheme must be established early, to guide the design process from top to bottom. One early design decision for our system is to have the ability to represent objects as a composition of components joined by non-rigid connections. This ability is desired for later experiments in class representation and learning. Incorporating this ability raises three major issues: 1) the representation scheme for primitive components; 2) the types of connection to be represented; and 3) how to interpret individual instances of class definition. Each of these three issues can be broken down into three topic areas: a) degree of realism; b) syntax of specification; and c) validity checking.

3.1. Representation of primitive components

A variety of well-known schemes are available for representing individual rigid components. Boundary surface descriptions have been around for some time and are still widely used. Generalized cylinders have become increasingly popular [Bro81a, Bro81b, Bro81c, Neva77, Ponc87]. Superquadrics have recently been proposed as another very general and flexible representation [Pent86]. Our choice of boundary surface description for primitive component definition should allow recognizable instances of most major subclasses in the class of objects chosen for study. The degree of realism should be adequate since our intended use of the system does not require exact fidelity to real-world instances of the objects. Our main interest is in modeling objects whose components may have a range of motion relative to a parameterized joint. We may later switch to the use of generalized cylinders to represent primitive components. The non-rigid parameterized connection type definitions should be able to remain unchanged when upgrading the representation system to allow non-planar component description.

An artifact is initially broken down into its components. These components can be identified through a natural breakdown of the artifact or can be a group of subcomponents whose relative relationship cannot change. Components can be entered individually in any orientation in the world coordinate system. It might be desired to input component descriptions into the system with the component center of mass aligned with the origin of the coordinate system, or it might be desired to enter components as they would be positioned in an instance of the object. The only stipulation is that all components must be entered relative to the same coordinate system. Further details of how components are defined and entered into the system are discussed in Section 4. Validity of a single component is decided by ensuring the component defined is a closed solid, with no dangling edges or planes [Requ80].

3.2. Types of connections to be represented

Before deciding what types of connection should be allowed in the system, it is useful to first consider the space of theoretically possible connection types. Connections can be divided into four categories, according to the relative movement allowed between components: 1) no relative movement; 2) rotational movement only; 3) translational movement only; and 4) a combination of translational and rotational movement. The first category, of course, would be a rigid connection between two components. Categories two and three have a small number of primitive possibilities, which can be combined to form the final category.

A connection allowing rotational movement can be characterized by the number of degrees of freedom allowed. Establishing one axis of rotation between components gives one degree of freedom. With only one degree of freedom the two components can be thought of as joined at an ideal hinge. Two axes of rotation allow two degrees of freedom and three establish three degrees of freedom. When two or three axes are established the joint between components can be thought of as a swivel joint or a ball joint, respectively. We require all axes of rotation for a single joint to be orthogonal and to intersect at a point. Definition of axes, establishing direction of rotation and determining order of rotation for the connection types chosen for this case study are discussed in Section 4.

The theoretically possible translational movements can be described in terms of: 1) the minimum required joint intersection between the two components, and 2) the space along which the minimum required joint intersection is allowed to vary. Imagine four different types of idealized joint intersections between two components (A and B): point, line segment, surface patch and volume. Table 1 describes the possible translational combinations of component A's joint intersection to component B. Due to the fact that the columns of Table 1 define a minimum joint intersection of component A to component B, the upper right area of the matrix is undefined. Component A's joint intersection can only vary along an area of its size or larger. It should be obvious that column one of Table 1 can be used to model any other joint intersection in the table by restricting the row parameters accordingly.

A point intersection, varying along a point in component B can be considered a rigid connection. A point intersection that can vary along a line segment could be defined to allow component A to slide along a slot in component B. This describes a one-degree-of-translation motion. To define two and three-degrees-of-translation motion, the point connection would be allowed to vary along a surface patch on or in component B or a volume contained within component B, respectively.

Table 1 Translational combinations of components. Degrees of freedom (D-o-F) specified for each join.

MINIMUM JOINT INTERSECTION / VARYING ALONG	POINT	LINE SEGMENT	SURFACE PATCH	VOLUME
POINT	valid (RIGID)			
LINE SEGMENT	valid 1 D-o-F	valid 1 D-o-F		
SURFACE PATCH	valid 2 D-o-F	valid 2 D-o-F	valid 2 D-o-F	
VOLUME	valid 3 D-o-F	valid 3 D-o-F	valid 3 D-o-F	valid 3 D-o-F

A line segment defined in component A can vary along a line segment, surface patch or volume within component B. This type of definition can more realistically model the concept of a slotted joint between components. A surface patch joint intersection allowed to vary along a surface patch can be used to model such things as a car on a road. A volume varying along a volume could be used to model such a thing as a fish in a fish tank. A fish can be modeled as a volume that can be positioned anywhere within the volume of the tank. Even though the fish is not physically attached to the water in the tank, the object in this scene could be thought of as an aquarium, where the tank and the fish are components of the aquarium and will not become disjoint (at least the fish hopes so).

To this point, we have described different primitive rotational and translational connections and possible ways of modeling them. Combinations of these allowed movements can easily be modeled by adhering to a predefined syntax which designates the order of movement. We now need to investigate what set, of all possible compositions of connection types, would be viable for the class of objects to be modeled here.

We are deriving possible connection types by first defining a set of fundamental connection types. These connection types will consist of a rigid connection type along with all possible connection types allowing one degree of freedom. This set would allow either a translational displacement between two components (vary along a line) or a rotational movement (single axis). Figure 1(a) helps to illustrate the different connection types and how they associate the components of the model. The chair depicted is a simple example that can be described as a barber chair. The chair can be raised in a telescoping manner. The rotary motion is about the axis which goes through the pedestal and base. The arms of the chair are rigidly attached to the seat. The back is allowed to recline, hinged to the seat of the chair. Connection types defined for this example chair include RIGID, SLIDE (telescoping action of pedestal), HINGED, and ROTARY. The RIGID type connection gives zero degrees of freedom between the components specified. SLIDE, HINGED, and ROTARY type connections permit one degree of freedom each.

This set of connection types is used to establish the primitive or fundamental types of connections necessary to define other complex connection types. A complex connection type is one which uses the composite of one or more connection types at the same joint. One connection that is suitable for this class of objects is SWIVEL, which allows two degrees of freedom. This connection type can be defined as a combination of the fundamental connection types HINGE and ROTARY. Complex connection types will have an explicit order of movement of the fundamental connection types used.

Parameters for each connection type are specific to information necessary for the connection. Initial input of an object can be thought of as an instance of an object model in its "home" position along with constraints on the variable parameters. The home position is defined as a zero displacement in any of the allowable degrees of freedom of any of the components. By entering the initial object model in the home position, ranges of motion can be specified by a single

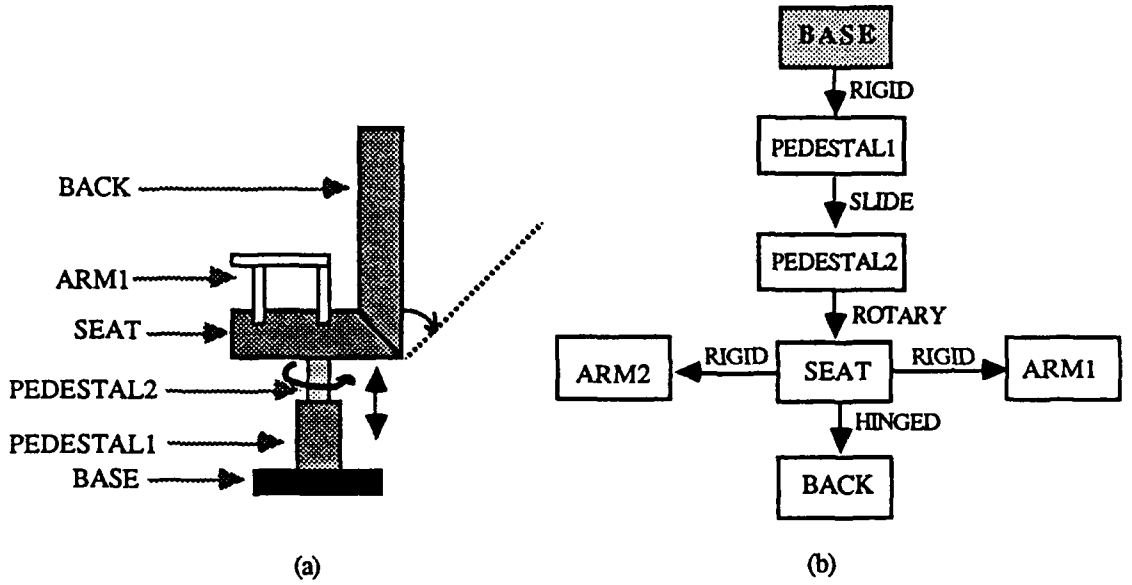


Figure 1 - (a) chair with labelled components and possible motion depicted
(b) directed graph structure of chair components with connection types

positive value rather than a range of negative to positive values. In this way motion is allowed from zero to some positive maximum displacement in distance or angular rotation.

Guaranteeing object validity starts by ensuring all components of the object are valid. The representation system "builds" objects by joining components in different manners. Validity of a connection between components is more of a conceptual question that needs to be answered. Obviously, two connected components cannot ever become totally disjoint and still be considered as having a valid connection. For two components to maintain a valid join there needs to be a minimum common connection. What we want is to allow a wide range for the degree of realism, and not force the user to represent objects at a lower level of detail than what they feel is necessary. Connection at a single point would not be used to represent objects realistically, but it is left as an option for the user to represent a rather non-realistic high level description of an object.

3.3. Interpretation of individual instance of class definition

An example of individual components of a chair, as it might naturally be broken down, is depicted in Figure 1(b). With connections depicted as arcs of a graph, parameterized connections are restricted such that the resulting graph must be acyclic (i.e. a tree). For example, the arms of the chair cannot be rigidly attached to the seat and the back at the same time. This cannot happen because of the allowable movement of the back relative to the seat. If the back of the chair was rigidly attached to the seat then these two components (seat and back) could be combined as a single component and then the arms could be joined to it.

The decision to restrict connections such that the graph structure formed is a connected graph with no cycles results in two main advantages. First the components of an object which allow movement cannot be joined in a manner that could result in an invalid configuration after movement. This type of error is depicted in the graph structure in Figure 2. Imagine component **A** is rigidly attached to component **B** and component **C**, yet component **B** is allowed to rotate relative to component **C**. It should be obvious that this is an unstable situation. Secondly, the tree structure can be used to connect components and propagate movement of the component parts. However, relative movement can affect more than one component. For example, consider the parameterized connection type labelled ROTARY between the pedestal and the seat of the chair depicted in Figure 1(a). When the seat rotates about the pedestal that same rotation must be applied to the arms and back of the chair. The same effect can be seen when the chair is raised (SLIDE type connection).

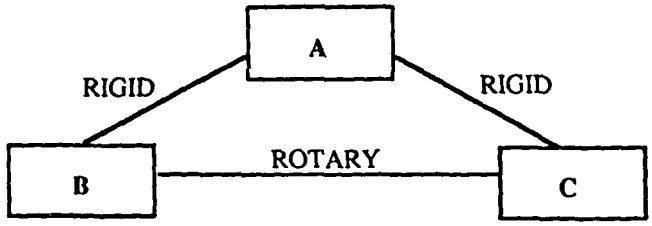


Figure 2 - Example of potentially unstable connection of components.

One of the major reasons for modeling 3-D objects is to be able to project 2-D images of the object in any orientation. Specifying an orientation is much the same as specifying a connection type ENVIRONMENT where one component is the coordinate system of the environment and the other component is the object. One component of the aggregate object must be chosen to set up the relationship of the object to its environment. We will refer to this component as BASE. This does not mean that the component chosen as BASE must be the one that rests on the ground as the actual base of the object. If the object rests on a single base, it might be easier to associate the orientation of the rest of the object to the base, but not necessary. One instance where this would not be the case would be found with the chair classified as a straight back chair which would be supported by four legs (Figure 3). It might be easier to assign the BASE attribute to the seat of the chair instead of one leg of the chair.

Once orientation of the object is set, all movement of parts is made about the BASE. For the chair example (Figure 1) the most logical selection of a base is the actual base of the chair. Once positioned, the base will not be affected by any parameterized connection type. The tree structure will be constructed with the BASE component as the root node. Arcs in the tree will have attributes of the connection type and any parameters necessary to specify the connection type. Propagation of movement progresses from the joint of connection through the entire subtree which has the connection component as its root in the connection graph. As the object is positioned, one component at a time, validity checks ensure that components that are not joined do not "collide" or interfere with one another. An instance of an object can then be displayed so that the user can visually validate the object's configuration. The degree of realism obtained for instantiations will be dependent on the degree of realism chosen when defining components and their connection types.

4. OBJECT REPRESENTATION SYSTEM

The component modeling technique is a planar face boundary representation; components are described by planar faces defined by their boundary edges and vertices. The system is broken down into three cooperating modules; the Object Editor Module, Object Instantiation Module, and Object Viewing Module.

4.1. Object Editor Module

The purpose of the Object Editor Module is to allow the user to enter the boundary surface description of the components of objects along with the connection type definitions that define the aggregate object.

4.1.1. Primitive component definition Objects are input to the system as a description of their components. Each component is defined by entering a list of faces, and each face is defined by a list of vertices in counterclockwise order. Face records are stored along with an array of vertices. As an individual component is input, described by its boundary surface description, component validity is checked. To be valid, the component must completely enclose a volume and not have any dangling lines or planes. The validity check of a component consists of the standard topological and geometric checks [Requ80]. As each face is entered, its vertices are checked for planarity. Verification that edges do not cross (two edges of a face meet at a vertex or not at all) is then completed. Once all faces have been input other checks can be accomplished. These checks include verification that each edge belongs to exactly two faces of the object and faces of an object intersect at an edge or not at all.

4.1.2. Connection type definitions The previously enumerated connection types emulate the major forms of component interconnections that actually occur in chairs, the class of objects chosen for this case study. Definition of the various connection types will require names of components and points to designate the orientation and location of the connection between the two components. These points will be known as *joint defining points*.The syntax of all connection type definitions requires eight essential parameters. Additional parameters will be required for non-rigid connection types. A template for connection type definitions can be written as:

Connection-Type-Name(ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, {*Optional Parameters*}).

The first two parameters designate the components being joined. Points P-*x*-A are joint defining points for ComponentA and points P-*x*-B for ComponentB. The three points of each component specify the coinciding *joint defining planes* for the two components. The first point on the first component will be made coincident with the first point of the second component, etc.. This gives an unambiguous alignment of the join.

Joint defining points for each component are added to the vertex list. All points that are used in connection type definitions of the object must be entered. In this way, if the component undergoes any transformations, its joint defining points will be transformed with it. These points are used in different manners, according to the connection type as described.

4.1.3. Specific connection types The most common fundamental connection type used will be RIGID. A RIGID connection indicates a fixed attachment allowing no relative movement between components. It should be obvious that the RIGID connection type would be redundant in the definition of any complex connection type. All complex connection types will therefore be a combination of fundamental connection types that permit one degree of freedom. To specify a RIGID connection requires only the eight essential parameters. A RIGID connection can be written as:

RIGID (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B).

An example of use of RIGID connection can be seen for the chair classified as straight back shown in Figure 3, in which all connections are of type RIGID. The components are defined as depicted in Figure 3(b). As can be seen, instances of components can be used in more than one orientation within the same object. The joint defining points can change for each instance definition. By using a component library, commonly used component shapes can be defined, differing only by a scale factor.

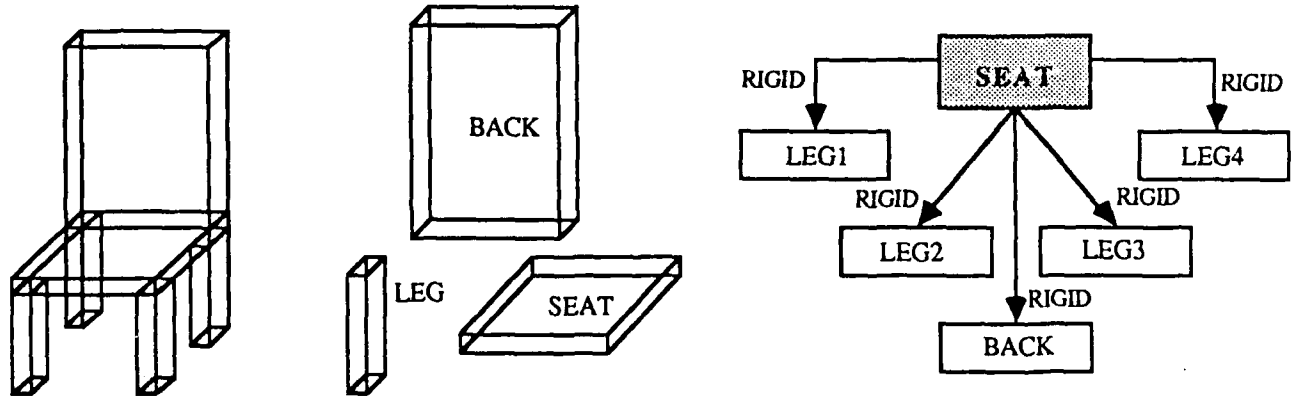


Figure 3 (a) Straight Back Chair (b) components of straight back chair (c) directed graph structure with connection types

Most representational systems allow only the RIGID type connection when defining objects by their components. This means that each instance of a swivel chair must be reinstantiated in the model base in the new orientation. It would be physically impossible to represent instances of all possible orientations of a chair that would be allowed to swivel about a base 360°. To allow such types of range of movement we have defined the following connection types: HINGED, ROTARY, SLIDE (telescoping action), and SWIVEL. These parameterized connections allow movement along or about an axis defined by the joint defining points.

To allow for a chair that can recline, the HINGED connection is defined as follows:

HINGED (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, Degrees).

A HINGED connection is specified by nine parameters. The first eight parameters are the essential parameters required for all connection definitions. The HINGED connection type differs from the RIGID type by allowing angular movement about the joint defining line formed by P-1-B and P-2-B. Connection of P-1-B to P-2-B sets up an axis of rotation for the HINGED connection. P-1-B is defined as the tail of the line directed toward P-2-B. By using the right hand rule with the implied direction of the line a positive rotational direction can be assigned. Parameter nine constrains the angular "hinge" movement allowed from the original home position. This parameter is assigned the value Degrees ($0° \leq \text{Degrees} < 360°$) when defining an object. This means that an instance of the object can be produced in an orientation where the angular displacement of 0 to Degrees can be instantiated between the two components. It will always be considered that ComponentB moves relative to ComponentA.

Another type of connection between components is ROTARY, defined as follows:

ROTARY (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, P-4-A, Degrees).

ROTARY requires ten parameters. As before, the first eight parameters correspond to the eight essential parameters. The second and third pair of parameters (parameters 3-6) define directed lines in the same manner as used in HINGED. Rotational movement is allowed, centered about P-1-A and P-1-B. Direction of angular rotation is defined by parameters P-1-A and P-4-A. Parameter nine, P-4-A, is a joint defining point found on a perpendicular to the joint defining plane containing P-1-A. A direction vector is defined with P-1-A as the tail and P-4-A the head. The vector (P-1-A,P-4-A) sets up an axis of rotation. Again, the right hand rule can be utilized to establish a positive rotational direction. Parameter ten constrains the angular ROTARY movement, and is assigned the value Degrees ($0° \leq \text{Degrees} < 360°$) when defining the object. Instantiation of the object is accomplished by choosing an angular displacement of 0 to Degrees. ComponentB will be displaced Degrees relative to ComponentA.

A SLIDE connection between components allows a translational displacement along an axis set up by the joint defining points rather than a rotational displacement. SLIDE connection is defined as follows:

SLIDE(ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, Distance).

The component names, coincident points and hence the joint defining planes are identified by the first eight essential parameters. Translational movement of the second component is allowed from the zero, or home, position to Distance displacement. This movement is allowed in the direction of the line defined in the first component (ComponentA) by the first pair of points associated to the first component (P-1-A to P-2-A sets up a direction vector). The only limitation on the Distance parameter is that the two components joined by the SLIDE connection type cannot be displaced to the point of making the components disjoint.

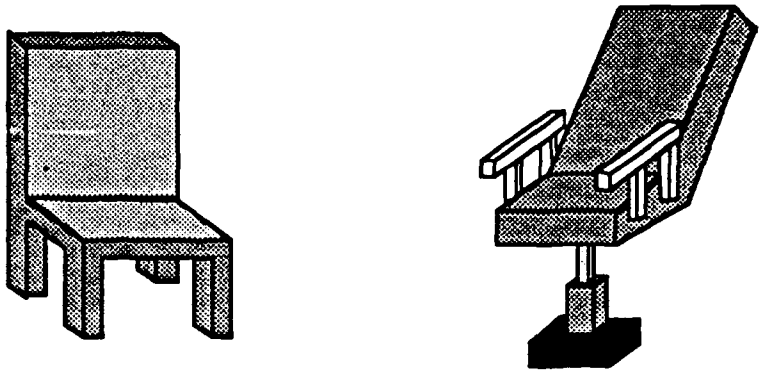
A SWIVEL connection is a complex connection type formed by a combination of the ROTARY and HINGED connection:

SWIVEL (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, P-4-A, Degrees1, Degrees2).

The first ten of the eleven parameters of the SWIVEL connection are directly related to the ten parameters of the ROTARY definition. The eleventh parameter is directly related to the ninth parameter of the HINGED connection relation. The order of angular displacement will be the rotational movement followed by the angular hinge movement. The rotary motion of the SWIVEL type connection keeps the first joint defining point of each component coincident and rotates the second component in a positive angular direction. Positive rotational direction and axis of rotation are set up by the directed vector P-1-A to P-4-A. It should be noted that we specify the axis of rotation for the hinged movement to be about the line contained in the second component (P-1-B,P-2-B). Therefore the axis of rotation of the hinged movement is rotated about the axis of rotation for the rotary motion. An example of this type of connection can be depicted by a simple office chair which can rotate about a base and recline.

4.2. Object Instantiation Module

The Object Instantiation Module will be used to create a representation of a specific configuration of a specific object. The instantiation module takes, as input, an object definition, a specific set of values for its connection parameters (if any), and orientation parameters, and creates, as its output, a file representing the surfaces of the object as it appears in the specific indicated configuration (Figure 4). This 3-D data can then be used as test data for later experiments dealing with recognition or class representation and learning.



(a) (b)

Figure 4 - Instantiations of (a) the straight back chair model and (b) the barber chair model.

4.3. Object Viewing Module

An Object Viewing Module is created to go with the instantiation module. The purpose of the viewing module is to input a particular object instantiation file and create projected views of the object as it would be seen from a representative set of viewpoints. This module is useful for checking out object instantiations to be sure that they represent the desired object and configuration. It could also be useful as test data for experiments in reconstructing 3-D shape from 2-D views.

5. SUMMARY AND FURTHER RESEARCH

The representation system described here is being implemented on a SUN workstation. Object instantiations created by the system will be used to create plausible test data that would constitute knowledge of physical form for the next stage of this project. Our future research plan, after completion of this representation system, is to combine the description of physical form with the knowledge of function to generalize a description of a class of objects. We feel this information can be used both to recognize the objects and also to learn new instances of objects in the class.

6. REFERENCES

Badl79 Badler, N.I. and O'Rourke, J., "Representation of Articulable, Quasi-Rigid, Three-Dimensional Objects," Presented at the NSF sponsored *Workshop on Three-Dimensional Object Representations*, (April, 1979).

Bied87 Biederman, Irving, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review*, Vol. 94, No. 2, (1987), 115-147.

Besl85 Besl, P.J. and Jain, R.C. "Three-Dimensional Object Recognition," *ACM Computing Surveys* 17, 1 (March 1985), 75-145.

Bro81a Brooks, R.A., and Binford, T.O. "Geometric Modeling in Vision for Manufacturing," *SPIE Techniques and Applications of Image Understanding*, 128, (1981), 141-159.

Bro81b Brooks, R.A. and Binford, T.O. "Representing and Reasoning About Partially Specified Scenes," *Proceedings of DARPA 1980 Image Understanding Workshop*, #SAI-81-170-WA (1981), 95-103.

Bro81c Brooks, R.A., "Symbolic Reasoning Among 3-D Models and 2-D Images," *Artificial Intelligence*, 17 (1981), 285-348.

Gold87 Goldberg, R.R. and Lowe, D.G., "Verification of 3-D Parametric Models in 2-D Image Data," *Proceedings of the IEEE Computer Society Workshop on Computer Vision*, (1987), 255-257.

Grim87 Grimson, W. and Eric L., "Recognition of Object Families Using Parameterized Models," *Proceedings First International Conference on Computer Vision*, (1987), 93-101.

Neva77 Nevatia, R. and Binford, T.O., "Description and Recognition of Curved Objects," *Artificial Intelligence* 8 (1977), 77-79.

Pent86 Pentland, A.P., "Perceptual Organization and the Representation of Natural Form," *Artificial Intelligence* 28 (1986), 293-331.

Ponc87 Ponce, J., Chelberg, D. Kriegman, D.J., and Mann, W., "Geometric Modelling with Generalized Cylinders", *Proceedings of the IEEE Computer Society Workshop on Computer Vision*, (1987), 268-270.

Requ80 Requicha, A.A.G., "Representations for Rigid Solids: Theory, Methods, and Systems," *ACM Computing Surveys* 12, 4 (1980), 437-464.

Requ82 Requicha, A.A.G., and Voelcker, H.B., "Solid Modeling: A Historical Summary and Contemporary Assessment," *IEEE Computer Graphics and Applications* (1982), 9-24.

PROCEEDINGS OF THE FIRST FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SYMPOSIUM

Orlando, Florida
Mar. 4-6, 1988

Mark B. Fishman
Editor

A 3-D REPRESENTATION SCHEME FOR SOLID OBJECTS WITH MOVEABLE COMPONENTS

Louise Stark and Kevin W. Bowyer
Department of Computer Science and Engineering
University of South Florida
Tampa, FL 33620

ABSTRACT

Few three-dimensional object representation systems allow non-rigid connections between components of the object. We define a representation scheme that permits parameterized non-rigid connections, allowing one definition to specify a range of permissible configurations of an aggregate object. This representation can be used to generate 3-D instantiations of particular configurations, and 2-D projected images of particular 3-D instantiations. The syntax of component description and connection types for our specific system is reviewed, along with the semantics of the allowable ranges of movement associated with each connection type. The actual representation system modules are also described.

1. INTRODUCTION

Descriptive representation of the physical form of 3-D solid objects is used in applications such as object recognition, CAD/CAM and graphics. Many different modeling techniques can be utilized to represent the object as a whole or to represent components (subparts) of the rigid object. Systems which allow objects to be defined as a connected structure of components must have a scheme to unambiguously specify the joining of components. Currently, most such systems allow only rigid connection between rigid components. We are investigating the use of non-rigid connections between rigid components.

We envision an interesting and versatile representation of physical form to be one which allows:

(1) definition of an "object" as a composition of multiple components, where
(2) one component may be attached to another component by any of several types of "connection" (each connection type defines a different type of possible relative movement between the components).

By allowing parameterized non-rigid connections, the system can represent an object which can take on any of a range of allowable configurations. Major issues that must be considered with this type of representation include: 1) the syntax of component and connection type definitions; 2) the semantics of the allowable ranges of movement associated with given types of connection; and 3) how to interpret shape under different configurations.

System design trade-offs must be made concerning such issues as ease of implementation, ease of use, degree of realism in representing components and connections, and breadth of class of objects representable. The system described here models components as three-dimensional planar face solids. This representation is easy to implement and relatively easy to use. The trade-off is that it is not as realistic as might be. However, we feel this representation is adequate to model object prototypes for experiments on class representation.

Background information is presented in Section 2 in the form of related literature. The system modules, along with the description of connection types and how they are defined, is covered in Section 3. A summary and recommendations for further research concludes the paper in Section 4.

2. RELATED WORK

Current literature on solid modeling deals mainly with rigid solid objects. Good surveys of the literature in this area can be found in (Besl and Jain 1985; Requicha 1980; Requicha and Voelcker 1982). Modeling of parameterized objects has been addressed by only a few researchers. The systems reviewed here incorporate the idea of objects whose components may have a range of orientations.

The use of generalized cylinders to model components of objects has become increasingly popular (Nevatia and Binford 1977; Brooks and Binford 1981a, 1981b; Brooks 1981). Aggregate objects are formed by joining the components in different manners. In (Nevatia and Binford 1977) some parts are allowed to be articulated, depending upon the type of joint defined. It is pointed out that, for the system described, "articulations of parts of an object are assumed to be completely unrestricted" and that without specified articulation limits, reliable discrimination between similar objects is difficult. ACRONYM (Brooks and Binford 1981a, 1981b; Brooks 1981) representation system allows modeling of specific instances, subclasses and classes of objects. Their representation allows for variations in size and structure between specific instances within a class.

The geometric modeler of the Successor (to ACRONYM) vision system uses a straight homogeneous generalized cylinder representation of components (Ponce et al. 1987). Affixments of components are specified "by spatial relationships between planar faces of primitives".

This work was supported by the AFOSR/UES Research Initiation Program under grant #S-760-00MG-003.

Proceedings of the 1st Florida Artificial Intelligence Research Symposium, 1988 by The Florida AI Research Symposium

Grimson deals with "families of objects that are characterized by sets of free parameters" (Grimson 1987). Object models are represented by sets of planar faces. The set of faces for a model is not restricted to be connected or complete. One class of objects demonstrated (scissors) allows a limited number of moving parts with a single degree of freedom. Stretching deformation was demonstrated for a family of hammers whose handles can stretch along the axis of the handle. The idea of multiple types of connection for a single object is not addressed.

(Badler and O'Rourke 1979) modeled objects that are representations of quasi-rigid segments connected at articulable point-like joints. The human body is modeled using "a representation for the object segments and joints, the surface and coordinate system of each segment, and well-defined mathematical relations between adjacent segment coordinate systems during movement". An instantiation of the body is established "with respect to the environment through a chain of special instances of joints and segments". Overlapping spheres are used to model the "skin" of the segments.

Pentland discusses modeling natural forms and artifacts in a "scene structure at a scale that is more like our naive perceptual notion of a 'part' " (Pentland 1986). The representation is based on superquadrics and fractals, which allows simple composition of components. The modeling primitives are associated to the "parts" of the form or artifact. Pentland suggests that a general vision system would follow the paradigm of a model base consisting of "parts that make up the specific object, rather than a model of the entire object, and the goal is to identify those component parts".

3. OBJECT REPRESENTATION SYSTEM

There are many important issues to be considered when designing a representation scheme, and trade-offs to be made with each design decision. Priorities concerning the type of information needed in the representation scheme must be established early, to guide the design process from top to bottom. One early design decision for our system is to have the ability to represent objects as a composition of components joined by non-rigid connections. This ability is desired for later experiments in class representation and learning. Incorporating this ability raises three major issues: 1) the representation scheme for primitive components; 2) the types of connection to be represented; and 3) how to interpret individual instances of class definition. These issues are addressed in the description of our representation system. The system is broken down into three cooperating modules; the Object Editor Module, Object Instantiation Module, and Object Viewing Module.

3.1 Object Editor Module

The purpose of the Object Editor Module is to allow the user to enter the boundary surface description of the components of objects along with the connection type definitions that define the aggregate object.

Our choice of boundary surface description for primitive component definition should allow recognizable instances of most major subclasses in the class of objects chosen for study. The degree of realism should be adequate since our intended use of the system does not require exact fidelity to real-world instances of the objects. Our main interest is in modeling objects whose components may have a range of motion relative to a parameterized joint. We may later switch to the use of generalized cylinders to represent primitive components. The non-rigid parameterized connection type definitions should be able to remain unchanged when upgrading the representation system to allow non-planar component description.

Primitive component definition An artifact is initially broken down into its components. These components can be identified through a natural breakdown of the artifact or can be a group of subcomponents whose relative relationship cannot change. Components can be entered individually in any orientation in the world coordinate system. It might be desired to input component descriptions into the system with the component center of mass aligned with the origin of the coordinate system, or it might be desired to enter components as they would be positioned in an instance of the object.

An example of individual components of a chair, as it might naturally be broken down, is depicted in Figure 1(b). With connections depicted as arcs of a graph, parameterized connections are restricted such that the resulting graph must be acyclic (i.e. a tree). For example, the arms of the chair cannot be rigidly attached to the seat and the back at the same time. This cannot happen because of the allowable movement of the back relative to the seat. If the back of the chair was rigidly attached to the seat then these two components (seat and back) could be combined as a single component and then the arms could be joined to it.

Each component is defined by entering a list of faces, and each face is defined by a list of vertices in counterclockwise order. Face records are stored along with an array of vertices. As an individual component is input, described by its boundary surface description, component validity is checked. To be valid, the component must completely enclose a volume and not have any dangling lines or planes. The validity check of a component consists of the standard topological and geometric checks (Requicha 1980).

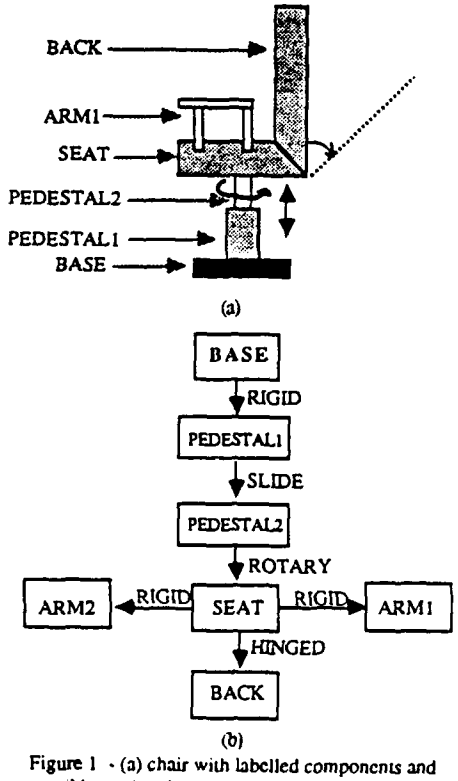


Figure 1 - (a) chair with labelled components and possible motion depicted (b) directed graph structure of chair components with connection types

Establishing Object Orientation One of the major reasons for modeling 3-D objects is to be able to project 2-D images of the object. Specifying an orientation is much the same as specifying a connection type ENVIRONMENT where one component is the coordinate system of the environment and the other component is the object. One component of the aggregate object must be chosen to set up the relationship of the object to its environment. We will refer to this component as BASE. This does not mean that the component chosen as BASE must be the one that rests on the ground as the actual base of the object. If the object rests on a single base, it might be easier to associate the orientation of the rest of the object to the base, but not necessary. One instance where this would not be the case would be found with the chair classified as a straight back chair which would be supported by four legs (Figure 2). It might be easier to assign the BASE attribute to the seat of the chair instead of one leg of the chair.

Once orientation of the object is set, all movement of parts is made about the BASE. For the chair example (Figure 1) the most logical selection of a base is the actual base of the chair. Once positioned, the base will not be affected by any parameterized connection type. The tree structure will be constructed with the BASE component as the root node. Arcs in the tree will have attributes of the connection type and any parameters necessary to specify the connection type. Propagation of movement progresses from the joint of connection through the entire subtree which has the connection component as its root in the connection graph. As the object is positioned, one component at a time, validity checks ensure that components that are not joined do not "collide" or interfere with one another. An instance of an object can then be displayed so that the user can visually validate the object's configuration. The degree of realism obtained for instantiations will be dependent on the degree of realism chosen when defining components and their connection types.

Connection type definitions Before deciding what types of connection should be allowed in the system, it is useful to first consider the space of theoretically possible connection types. Connections can be divided into four categories, according to the relative movement allowed between components: 1) no relative movement; 2) rotational movement only; 3) translational movement only; and 4) a combination of translational and rotational movement. The first category, of course, would be a rigid connection between two components. Categories two and three have a small number of primitive possibilities, which can be combined to form the final category. A more thorough explanation of the considerations underlying possible connection types appears in (Stark and Bowyer 1988).

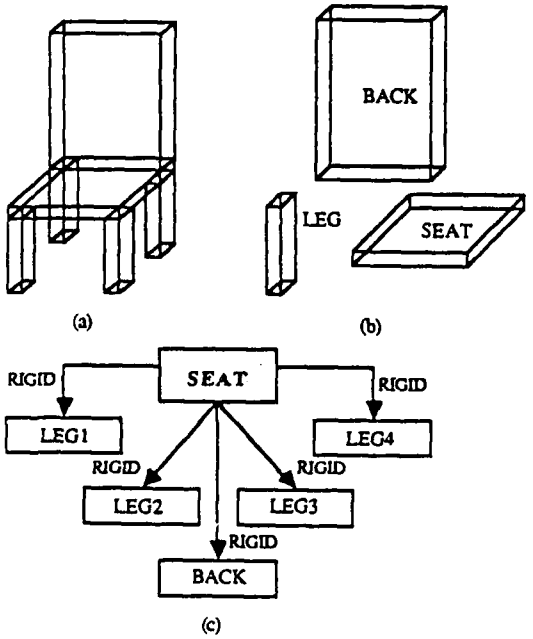


Figure 2 - (a) Straight Back Chair (b) components of straight back chair (c) directed graph structure with connection types

We are deriving possible connection types by first defining a set of fundamental connection types. These connection types will consist of a rigid connection type along with all possible connection types allowing one degree of freedom. This set would allow either a translational displacement between two components or a rotational movement. Figure 1(a) helps to illustrate the different connection types and how they associate the components of the model. The chair depicted is a simple example that can be described as a barber chair. The chair can be raised in a telescoping manner. The rotary motion is about the axis which goes through the pedestal and base. The arms of the chair are rigidly attached to the seat. The back is allowed to recline, hinged to the seat of the chair. Connection types defined for this example chair include RIGID, SLIDE (telescoping action of pedestal), HINGED, and ROTARY. The RIGID type connection gives zero degrees of freedom between the components specified. SLIDE, HINGED, and ROTARY type connections permit one degree of freedom each.

This set of connection types is used to establish the primitive or fundamental types of connections necessary to define other complex connection types. A complex connection type is one which uses the composite of one or more connection types at the same joint. One connection that is suitable for this class of objects is SWIVEL, which allows two degrees of freedom. This connection type can be defined as a combination of the fundamental connection types HINGE and ROTARY. Complex connection types will have an explicit order of movement of the fundamental connection types used.

The previously enumerated connection types emulate the major forms of component interconnections that actually occur in chairs, the class of objects chosen for this case study. Definition of the various connection types will require names of components and points to designate the orientation and location of the connection between the two components. These points will be known as *joint defining points*. The syntax of all connection type definitions requires eight essential parameters. Additional parameters will be required for non-rigid connection types. A template for connection type definitions can be written as:

Connection-Type(ComponentA, ComponentB,
P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B,
{*Optional Parameters*}).

The first two parameters designate the components being joined. Points P-*x*-A are joint defining points for ComponentA and points P-*x*-B for ComponentB. The three points of each component specify the coinciding *joint defining planes* for the two components. The first point on the first component will be made coincident with the first point of the second component, etc.. This gives an unambiguous alignment of the join.

Joint defining points for each component are added to the vertex list. All points that are used in connection type definitions of the object must be entered. In this way, if the component undergoes any transformations, its joint defining points will be transformed with it. These points are used in different manners, according to the connection type as described.

Parameters for each connection type are specific to information necessary for the connection. Initial input of an object can be thought of as an instance of an object model in its "home" position, along with constraints on the variable parameters. The home position is defined as a zero displacement in each of the allowable degrees of freedom of any of the components. By entering the initial object model in the home position, ranges of motion can be specified by a single positive value rather than a range of negative to positive values. In this way motion is allowed from zero to some positive maximum displacement in distance or angular rotation.

Specific connection types The most common fundamental connection type used will be RIGID. A RIGID connection indicates a fixed attachment allowing no relative movement between components. It should be obvious that the RIGID connection type would be redundant in the definition of any complex connection type. All complex connection types will therefore be a combination of fundamental connection types that permit one degree of freedom. To specify a RIGID connection requires only the eight essential parameters. A RIGID connection can be written as:

RIGID (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B).

An example of use of RIGID connection can be seen for the chair classified as straight back shown in Figure 2, in which all connections are of type RIGID. The components are defined as depicted in Figure 2(b). As can be seen, instances of components can be used in more than one orientation within the same object. The joint defining points can change for each instance definition. By using a component library, common component shapes can be defined, differing only by a scale factor.

Most representational systems allow only the RIGID type connection when defining objects by their components. This means that each instance of a swivel chair must be reinstantiated in the model base in the new orientation. It would be physically impossible to represent instances of all possible orientations of a chair that would be allowed to swivel about a base 360°. To allow such types of range of movement we have defined the following connection types: HINGED, ROTARY, SLIDE (telescoping action), and SWIVEL. These parameterized connections allow movement along or about an axis defined by the joint defining points.

To allow for a chair that can recline, the HINGED connection is defined as follows:

HINGED (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, Degrees).

A HINGED connection is specified by nine parameters. The first eight parameters are the essential parameters required for all connection definitions. The HINGED connection type differs from the RIGID type by allowing angular movement about the joint defining line formed by P-1-B and P-2-B. Connection of P-1-B to P-2-B sets up an axis of rotation for the HINGED connection. P-1-B is defined as the tail of the line directed toward P-2-B. By using the right hand rule with the implied direction of the line a positive rotational direction can be assigned. Parameter nine constrains the angular "hinge" movement allowed from the original home position. This parameter is assigned the value Degrees ($0° \leq$ Degrees $< 360°$) when defining an object. This means that an instance of the object can be produced in an orientation where the angular displacement of 0 to Degrees can be instantiated between the two components. It will always be considered that ComponentB moves relative to ComponentA.

Another type of connection between components is ROTARY, defined as follows:

ROTARY(ComponentA,ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B,P-4-A, Degrees).

ROTARY requires ten parameters. As before, the first eight parameters correspond to the eight essential parameters. The second and third pair of parameters (parameters 3-6) define directed lines in the same manner as used in HINGED. Rotational movement is allowed, centered about P-1-A and P-1-B. Direction of angular rotation is defined by parameters P-1-A and P-4-A. Parameter nine, P-4-A, is a joint defining point found on a perpendicular to the joint defining plane containing P-1-A. A direction vector is defined with P-1-A as the tail and P-4-A the head. The vector (P-1-A,P-4-A) sets up an axis of rotation. Again, the right hand rule can be utilized to establish a positive rotational direction. Parameter ten constrains the angular ROTARY movement, and is assigned the value Degrees ($0° \leq$ Degrees $< 360°$) when defining the object. Instantiation of the object is accomplished by choosing an angular displacement of 0 to Degrees. ComponentB will be displaced Degrees relative to ComponentA.

A SLIDE connection between components allows a translational displacement along an axis set up by the joint defining points rather than a rotational displacement. SLIDE connection is defined as follows:

SLIDE(ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, Distance).

The component names, coincident points and hence the joint defining planes are identified by the first eight essential parameters. Translational movement of the second component is allowed from the zero, or home, position to Distance displacement. This movement is allowed in the direction of the line defined in the first component (ComponentA) by the first pair of points associated to the first component (P-1-A to P-2-A sets up a direction vector). The only limitation on the Distance parameter is that the two components joined by the SLIDE connection type cannot be displaced to the point of making the components disjoint.

A SWIVEL connection is a complex connection type formed by a combination of the ROTARY and HINGED connection:

SWIVEL (ComponentA, ComponentB, P-1-A, P-1-B, P-2-A, P-2-B, P-3-A, P-3-B, P-4-A, Degrees1, Degrees2).

The first ten of the eleven parameters of the SWIVEL connection are directly related to the ten parameters of the ROTARY definition. The eleventh parameter is directly related to the ninth parameter of the HINGED connection relation. The order of angular displacement will be the rotational movement followed by the angular hinge movement. The rotary motion of the SWIVEL type

connection keeps the first joint defining point of each component coincident and rotates the second component in a positive angular direction. Positive rotational direction and axis of rotation are set up by the directed vector P-1-A to P-4-A. It should be noted that we specify the axis of rotation for the hinged movement to be about the line contained in the second component (P-1-B,P-2-B). Therefore the axis of rotation of the hinged movement is rotated about the axis of rotation for the rotary motion. An example of this type of connection can be depicted by a simple office chair which can rotate about a base and recline.

3.2 Object Instantiation Module

The Object Instantiation Module will be used to create a representation of a specific configuration of a specific object. The instantiation module takes, as input, an object definition, a specific set of values for its connection parameters (if any), and orientation parameters, and creates, as its output, a file representing the surfaces of the object as it appears in the specific indicated configuration (Figure 3). This 3-D data can then be used as test data for later experiments dealing with recognition or class representation and learning.

3.2 Object Viewing Module

An Object Viewing Module is created to go with the instantiation module. The purpose of the viewing module is to input a particular object instantiation file and create projected views of the object as it would be seen from a representative set of viewpoints. This module is useful for checking out object instantiations to be sure that they represent the desired object and configuration. It could also be useful as test data for experiments in reconstructing 3-D shape from 2-D views.

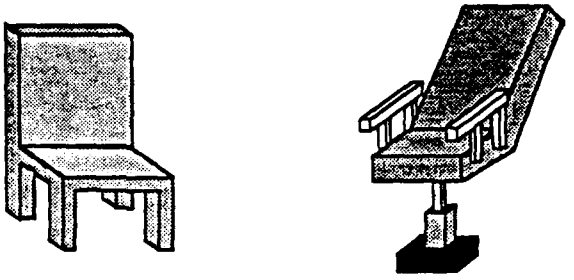


Figure 3 - Instantiations of (a) the straight back chair model and (b) the barber chair model.

4. SUMMARY AND FURTHER RESEARCH

The representation system described here is being implemented on a SUN workstation. Object instantiations created by the system will be used to create plausible test data that would constitute knowledge of physical form for the next stage of this project. Our future research plan is to combine the description of physical form with the knowledge of function to generalize a description of a class of objects. We feel this information can be used both to recognize the objects and also to learn new instances of objects.

5. REFERENCES

Badler, N.I. and O'Rourke, J., "Representation of Articulable, Quasi-Rigid, Three-Dimensional Objects," Presented at the NSF sponsored *Workshop on Three-Dimensional Object Representations*, (April, 1979).

Besl, P.J. and Jain, R.C. "Three-Dimensional Object Recognition," *ACM Computing Surveys* 17, 1 (March 1985), 75-145.

Brooks, R.A., and Binford, T.O. "Geometric Modeling in Vision for Manufacturing," *SPIE Techniques and Applications of Image Understanding*, 128, (1981), 141-159.

Brooks, R.A. and Binford, T.O. "Representing and Reasoning About Partially Specified Scenes," *Proceedings of DARPA 1980 Image Understanding Workshop*, #SAI-81-170-WA (1981), 95-103.

Brooks, R.A., "Symbolic Reasoning Among 3-D Models and 2-D Images," *Artificial Intelligence*, 17 (1981), 285-348.

Grimson, W. Eric L., "Recognition of Object Families Using Parameterized Models," *Proceedings First International Conference on Computer Vision*, (1987), 93-101.

Nevatia, R. and Binford, T.O., "Description and Recognition of Curved Objects," *Artificial Intelligence* 8 (1977), 77-79.

Pentland, A.P., "Perceptual Organization and the Representation of Natural Form," *Artificial Intelligence* 28 (1986), 293-331.

Ponce, J., Chelberg, D. Kriegman, D.J., and Mann, W., "Geometric Modelling with Generalized Cylinders", *Proceedings of the IEEE Computer Society Workshop on Computer Vision*, (1987), 268-270.

Requicha, A.A.G., "Representations for Rigid Solids: Theory, Methods, and Systems," *ACM Computing Surveys* 12, 4 (1980), 437-464.

Requicha, A.A.G., and Voelcker, H.B., "Solid Modeling: A Historical Summary and Contemporary Assessment," *IEEE Computer Graphics and Applications* (1982), 9-24.

Stark, L. and Bowyer, K., "Representation of 3-D Objects Using Non-Rigid Connection of Components," *SPIE 1988 Technical Symposium Southeast on Optics, Electro-Optics and Sensors*, Orlando, FL, (4-8 April 1988).

FINAL REPORT

DEVELOPMENT AND EVALUATION OF A BAYESIAN TEST FOR SYSTEM TESTABILITY

by

Ronald V. Canfield
Utah State University
Logan, Utah 84322-3900

27 DECEMBER 1988

ABSTRACT

Bayesian testability demonstration plans are developed in this report. Existing design criteria based on average risks is found to produce designs which are inconsistent with the logical effect of prior information. A Bayes modified minimax risk criterion is developed to correct this problem. The Bayes modified minimax approach is used to derive sample tables for design and to illustrate the method of selecting an appropriate experimental design. The sensitivity of designs based on this risk criteria is investigated. The method is shown to be very robust. Bayes modified minimax designs are compared with classical fixed sample designs to illustrate the effect of prior information.

1. INTRODUCTION

Technological advances have greatly increased the capability and complexity of modern electronic systems. Maintenance and readiness requirements of such systems has necessitated the use of automatic diagnostic equipment. Specification of automatic diagnostic capability includes testability requirements, e.g., 90% of all possible faults will be isolated by the diagnostic system. Contracts may also specify that the producer demonstrate compliance via a statistical test called a testability demonstration. This report documents the effort under U.S. Air Force contract to develop Bayesian testability demonstration plans.

Sections 2 and 3 of this report are a general review of existing experimental design principles for Bayesian tests. The existing risk criteria are examined and found to have a serious logical weakness. A new risk criterion based on a modification of the minimax strategy is developed in section 3. This new risk criterion considerably simplifies the problem quantifying prior information. Sections 4-8 are specific to the problem of testability demonstration. The loss functions, prior distributions and test format for this application are derived. Comparison of this Bayesian test with the classical test is made to assist in the interpretation of risk in the Bayesian case. Sensitivity of the Bayesian test to improper prior distributions is examined.

A sequential test for testability demonstration was investigated. Since a sequential test uses the posterior distribution the new risk criterion does not apply. Therefore the simplification benefits only the fixed sample test. The sequential problem remains very difficult (Berger, 1985.) Administration of existing sequential tests by personnel with a great deal of statistical training will be difficult. Given the usually limited background of most individuals involved with the task of carrying out testability demonstration, the sequential test as it now exits seems to have limited utility. For this reason this report considers only the fixed sample test.

1.1 Project Goals

The primary objective of this project is to develop a useful fixed sample test for testability demonstration. To be useful, the test must be consistent with the knowledge and capabilities of personnel likely to be involved with test administration. The prior and cost information requirements must also be realistic.

A basic decision theory approach is used. By going back to the basics, a second look at some of the existing methods reveals deficiencies. A new risk criterion is introduced to correct the problems inherent with existing methods.

2. TEST ENVIRONMENT AND LITERATURE REVIEW

Bayes methods are attractive because they offer the potential of reducing dependence on sample information by using existing or prior information in the decision process. This additional information can reduce the expense of sampling without sacrificing a preset standard of acceptable risk. To be useful a test must be compatible with the personnel and information environment within which it is applied. This environment is examined so that the new test will be consistent with available prior information. An effort is also made to simplify the choice of prior distribution and loss functions so that the Bayesian test is a relatively simple alternative to the classical tests. This consideration is very important if the potential savings resulting from the use of prior information is to be realized. Much of the existing literature on Bayesian theory is devoted to efficient use of information at the expense of test complexity. The most efficient test is of little value if so complicated that it is seldom used.

2.1. Test Environment

In order to develop a useful test it is important to understand the background of potential users of the test and the prior information which is typically available. In many cases the personnel who administer these tests are not statistical specialists. Thus it is important to minimize test complexity (e.g., Heiser, 1986.) However these individuals have very likely encountered classical tests in the past. Classical test designs require judgment with regard to choice of error rates, e.g., $\alpha = 0.05$. These judgments are made routinely so that experience provides an intuitive grasp of the effects of decision errors in terms of probabilities.

Bayesian methods add a new dimension to statistical tests, prior information. In most situations prior information is subjective or based on past performance of the producer on similar products. Quantification of prior information is not a frequent task of most analysts. Therefore sufficient opportunity to develop a strong intuitive understanding of the process is usually lacking. For this reason analysts are counseled to be conservative. This is usually interpreted to mean that a broad or relatively flat prior density is appropriate. It is shown in section 5 of this report that this perception is false when using existing experimental design methods.

A very useful concept in the decision theory approach to Bayesian analysis is the loss function. Although the concept has intuitive appeal, it requires cost information with regard to sampling expense and decision error losses. This information is often difficult to obtain especially when errors may involve loss of life. In addition to this difficulty, the use of real costs makes each test unique. Thus more complication is introduced by effectively eliminating the use of general tables for test plans. For these reasons many test plans (in effect) are based upon a simple 0-1 loss function. This approach ignores decision error effects due to parameter values in a region likely to occur in

most tests. Simple loss functions which have general utility are presented in this report.

2.2 Literature Review

There is a great deal of literature on Bayesian experimental design. The effort here is to review standard approaches in both theoretical and applications literature. Particular attention is given to the risk criteria which is used.

Experimental design for the decision problem $H_o: \theta \leq \theta_o$ against the alternate $H_1: \theta > \theta_o$ is considered in this section. The term preposterior analysis is used in Bayesian literature for this activity. Let $L(\theta, \delta(\underline{X}))$ represent the loss function for using decision $\delta(\underline{X})$ when the sample vector $\underline{X}$ of length n is observed and the parameter has value θ. The risk function is defined

$$R(\theta,\delta) = E_{\underline{X}}(L(\theta, \delta(\underline{X})) \tag{2-1}$$

and the Bayes risk is given by

$$r(\delta) = E_{\theta}(R(\theta,\delta)). \tag{2-2}$$

Let $h(\theta)$ represent the prior density function of θ. The appropriate decision function $\delta(\underline{X})$ (if it exists) minimizes (2-2). A theoretically satisfying approach to determine sample size is given by Berger (1985). This method includes sampling expense in the loss function. The optimal sample size for a test minimizes (2-2).

In most applications involving military contracts, opposing interests of the producer and the consumer demand that individual protection against decision errors be given. Under this obligation the Bayes risk is separated into the two components

$$r(\delta) = \int_{\theta>\theta_o}\int_{\underline{X}:\delta(\underline{X})=a_0} L(\theta,a_0) f(\underline{X}|\theta) h(\theta)\, d\underline{X}\, d\theta + \int_{\theta\leq\theta_o}\int_{\underline{X}:\delta(\underline{X})=a_1} L(\theta,a_1) f(\underline{X}|\theta) h(\theta)\, d\underline{X}\, d\theta \tag{2-3}$$

where a_0 represents the action "accept H_o" and a_1 "reject H_o". The first component on the right of (2-3) is referred to as the consumer risk and the second component, producer risk. Sample size is determined to satisfy acceptable constraints on the risk components in (2-3) (e.g., Martz and Waller, 1982.)

In most applications a parameter value $\theta_1 < \theta_o$ is chosen such that serious loss will result if $\theta < \theta_1$ and H_o is accepted. Although a loss function is usually not mentioned with this risk format the approach is equivalent to using the 0-1 loss function

$$L(\theta,a_0) = \begin{cases} 0 \text{ if } \theta < \theta_1 \\ 1 \text{ if } \theta \geq \theta_1 \end{cases} \tag{2-4}$$

and

$$L(\theta,a_1) = \begin{cases} 0 \text{ if } \theta > \theta_o \\ 1 \text{ if } \theta \leq \theta_o \end{cases} \tag{2-5}$$

This risk format (suggested by Easterling (1970) is referred to as average Bayes risk. Two aspects of this format are important to review because they will be addressed in a later section of this report. First, risk is computed as an average over all possible values of θ. Second the loss function ignores the losses which occur if $\theta_1 < \theta < \theta_0$.

The posterior risk (Schick and Drnas, 1972) is an alternative to the previous risk. It is very similar in mathematical representation. It is not necessary in this report to consider the precise formula for this risk. It suffices here to refer to the work of Goel and Joglekar (1976) where it is shown that the consumer and producer posterior Bayes risks are proportional to the average Bayes risks. Thus they also are the result of an averaging process.

3. RISK CRITERIA

The risk criteria described briefly in section 2 is examined more closely is this section. The rich mathematical framework of Bayesian analysis provides many possible definitions of risk. Choice of risk criteria is a important aspect of this method of analysis. The average Bayes risk and the posterior Bayes risk seem to be the most used in Air Force application. A major result of this study is the exposure of a basic defect in the response of fixed sample test plans based on these risks to changes in prior information content.

3.1 Average Risks

Berger (1985) gives a compelling argument for the usefulness of "rationality and coherency" developments as powerful tools in the effort to justify the Bayesian approach and to expose the "irrationality of purported truths in statistics." The same approach is used here to demonstrate "irrationality" of preposterior analysis using average risk criteria as in (2-2) and (2-3). It is shown in this section that the risk criteria rewards prior ignorance. Presumably if an organization invests the time and expense of developing more prior information, that investment should result in a decreased demand for sample information. The term "prior information" as used in this paper is an intuitive concept. Precise definitions in terms of entropy or Fisher's information are possible (e.g., Jaynes, 1968 and Canfield, 1977) but not necessary for purpose of this report. Intuitively if two priors have the same mean, then the one with the smaller variance has more prior information. For both approaches to preposterior analysis it is shown that an increase in prior information (at the same mean) can result in an increase in required sample size.

Consider two normally distributed priors $h_1(\theta) \sim N(\mu, \sigma_1^2)$ and $h_2(\theta) \sim N(\mu, \sigma_2^2)$ with $\sigma_1 < \sigma_2$ and identical means. There is a natural preference ordering of priors, i.e., $h_1(\theta)$ is preferred to $h_2(\theta)$ because it has more information. Preposterior analysis also imposes a preference ordering on priors, i.e., the prior requiring the smaller sample size is preferred. In this case effect of the prior on sample size is due to amount of information, not location of θ. Intuitively the amount of information in a prior is related to the influence of the prior in Bayesian analysis. It seems rational to expect that the above preference orderings should agree. This is not necessarily the case using average risk criteria.

Consider the following example taken from Berger (1985, example 3, case 2, page 438). Let the vector random variable $\underline{X}$ represent a random sample of size n taken from a $N(\theta, \sigma^2)$ population where σ is known. It is desired to test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Let $\delta(\underline{X}) = a_0$ represent the decision "accept H_0" and let $\delta(\underline{X}) = a_1$ represent the decision "reject H_0." The loss function for this test is given by

$$L(\ ,\ (X\ )) = \begin{cases} (\theta - \theta_0)^2 + nc & \text{for } \delta(\underline{X}) = a_0 \text{ and } \theta > \theta_0 \\ (\theta - \theta_0)^2 + nc & \text{for } \delta(\underline{X}) = a_1 \text{ and } \theta < \theta_0 \\ 0 \text{ otherwise,} \end{cases} \quad (3\text{-}1)$$

where n is the sample size and c is the cost of sampling per experimental unit. The prior density on θ is $N(\mu, \tau^2)$. The sample size which minimizes the overall risk is shown by Berger (1985) to be of the form

$$n(\tau) = A \exp(-B/\tau^2)/\tau \quad (3\text{-}2)$$

where A and B are positive constants independent of the prior variance τ^2. Note that maximum $n(\tau)$ occurs at $\tau = 2B$. For large prior variance ($\tau > 2B$), $n(\tau)$ is a decreasing function of τ. Thus large variance results in small sample size. Large prior variance is generally associated with conservatism in prior selection.

A similar relationship with prior information can be observed when consumer and producer risks are independently controlled as in (2-3) and for both average and posterior Bayes risk criteria. Consider sample size determination for a test of the mean of a negative exponential random variable using an inverted gamma prior having mean μ and shape parameter λ. Sampling plans for this test have been published by Goel and Joglekar (1976). The entries in table 1 are taken from Goel and Joglekar (1976). The variance of an inverted gamma random variable is a monotone decreasing function of λ. Note that the design sample size (time on test) can increase with a decrease in prior variance. This response is contrary to rational anticipation.

Table 3-1. Required sample size for the same producer and consumer risks (average and posterior Bayes risk).

		Sample Size (time on test) Risk Criteria	
		Average	Posterior
71.25	2	223	127
71.25	4	319	171
71.25	6	396	

Analysts are cautioned to be conservative when quantifying prior information since prior information is rarely precise. This is the primary motivation for the maximum entropy prior. Berger (1985) uses this argument to defend the natural conservatism of Bayesian methods as compared with the extreme conservatism of minimax methods. It is evident from the previous examples that the notions of conservatism in Bayesian estimation do not extend to experimental design.

An explanation for this conflicting response lies in interpretation of the prior distribution. In most situations even if the parameter can be legitimately regarded as a random variable, the prior distribution represents the present state of ignorance with respect to the true density. There is usually no really

satisfying method with convergence or bias criteria, etc. for estimating the prior density. This view is in contrast with the notion that θ is a random variable with a unique or "true" density which is estimable. The only safe approach is a conservative one in this case. However a conservative approach destroys the frequency interpretation of probability with respect to θ.

It may seem unnecessary to draw this distinction as it is clearly implied in most discussions of the prior distribution. However in this work it is important because irrationality of preposterior analysis has its roots in the interpretation of prior information. If the prior is a distribution of a true random variable, then a large prior variance may easily provide information for a reduction in risk. This is because a large variance increases the probability that the parameter will be in the extremes of its range. A small sample can easily detect values of the parameter at its extremes. On the other hand a small variance may concentrate probability in a range where the sample information is less efficient in detecting the position of the parameter. Both risk criteria average the decision risk to obtain the Bayes risks used to determine sample size. In reality a broad flat prior does not mean that the values of the parameter occur more frequently in the tails. It simply means that prior information is not precise. Thus contrary to popular perception, conservatism in experimental design cannot be achieved "naturally" using a broad, flat prior density. A risk strategy discussed in the next section uses prior information to modify the "over kill" of minimax strategy to provide a more credible risk criteria.

3.2 Bayes Modified Minimax Risk

The previous discussion suggests that the prior may not provide the means of addressing conservatism in experimental design. This is more appropriately accomplished by consideration of the perceived risk. Consider Bayes risk (2-2) and let $\delta(\underline{X})$ be the Bayes decision function if it exists.

The function of prior information is to provide a more realistic perception of risk by introducing "weighting" (i.e., the prior density $h(\theta)$) for the possible values of θ. The Bayes modified maximum risk $R^*(\theta,\delta)$ is defined

$$R^*(\theta,\delta) = \begin{cases} R(\theta_0^*,\delta) & \text{for } \theta \geq \theta_0 \\ R(\theta_1^*,\delta) & \text{for } \theta < \theta_0 \end{cases} \qquad (3\text{-}3)$$

where $\theta_0^* \geq \theta_0$ and $\theta_1^* < \theta_0$ are the values which maximize the total of the weighted risks associated with the two decision errors. That is, θ_0^* and θ_1^* are chosen such that

$$\underset{\theta \geq \theta_0}{h(\theta)R(\theta,\delta)} + \underset{\theta < \theta_0}{h(\theta)R(\theta,\delta)} \leq \underset{\theta_0^* \geq \theta_0}{h(\theta_0^*)R(\theta_0^*,\delta)} + \underset{\theta_1^* < \theta_0}{h(\theta_1^*)R(\theta_1^*,\delta)}. \qquad (3\text{-}4)$$

The parameter values θ_0^* and θ_1^* represent the positions of maximum "credible" risk for the decisions a_0 and a_1 respectively. The use of total weighted risk in (6) rather than individual maximization of the risks associated with the two decision errors is motivated by mathematical convenience. The optimal sample

size may depend depend on θ_0^* and θ_1^* jointly. This measure of risk has the intuitively appealing property that if the prior is flat (little prior information) the values θ_0^* and θ_1^* approach the respective values which yield the absolute maximums of $R(\theta,\delta)$ in H_0 and H_1 respectively. Given a more informative prior the values θ_0^* and θ_1^* will tend toward the mode of $h(\theta)$, i.e., the prior will have more influence. The Bayes modified minimax risk is defined:

$$r^*(\delta) = E_\theta(R^*(\theta,\delta)) = R(\theta_0^*,\delta)P(\theta \geq \theta_0) + R(\theta_1^*,\delta)P(\theta < \theta_0) \quad (3\text{-}5)$$

Sample size n and critical region for the Bayes rule are chosen to minimize (3-5) if sampling expense is considered. If consumer/producer risks are to be controlled, the critical region and sample size are used to bound the risk components on the right side of (3-5).

3.2.1 Bayes modified minimax risk example. Let $\underline{X}$ represent the sample vector of n independent normal random variables with identical but unknown mean θ and known standard deviation 5. The hypothesis $H_0: \theta \leq 10$ vs. $H_1: \theta > 10$ is to be tested. The loss function is given by

$$L(\theta,\delta(\underline{X})) = \begin{cases} 1000(\theta - \theta_0)^2 + n & \text{for } \delta(\underline{X}) = a_0 \text{ and } \theta > 10 \\ 1000(\theta - \theta_0)^2 + n & \text{for } \delta(\underline{X}) = a_1 \text{ and } \theta \leq 10 \\ 0 \text{ otherwise.} \end{cases} \quad (3\text{-}6)$$

Equation (3-6) indicates that the cost of sampling is one cost unit per experimental unit. Let the prior density of θ be normal with mean 11 and standard deviation σ. The Bayes decision rule for this example (given by Berger, 1985) is to take action a_0 if $\bar{X} < K$ and a_1 if $\bar{X} \geq K$. The Bayes minimax risk (3-5) for this example is

$$r^*(\delta) = 1000((\theta_0^* - \theta_0)^2 P(\bar{X} \geq K | \theta_0^*) P_0 + (\theta_1^* - \theta_0)^2 P(\bar{X} < K | \theta_1^*) P_1 \quad (3\text{-}7)$$

where $P_0 = P(\theta \leq 10)$ and $P_1 = P(\theta > 10)$. The value of K which minimizes (3-7) is easily found by taking the derivative of $r^*(\delta)$ with respect to K, setting to 0 and solving for K. For this case

$$K = (2\sigma^2(\ln((\theta_0^* - \theta_0)^2 P_0) - \ln((\theta_1^* - \theta_0) P_1))/n + \theta_0^{*2} - \theta_1^{*2})/2(\theta_0^* - \theta_1^*)).$$

The values of θ_0^* and θ_1^* are determined by numerically maximizing the total risk (3-5). Figure 3-1 shows a plot of design sample size vs. prior standard deviation σ. As expected the design sample size is a monotone increasing function of σ.

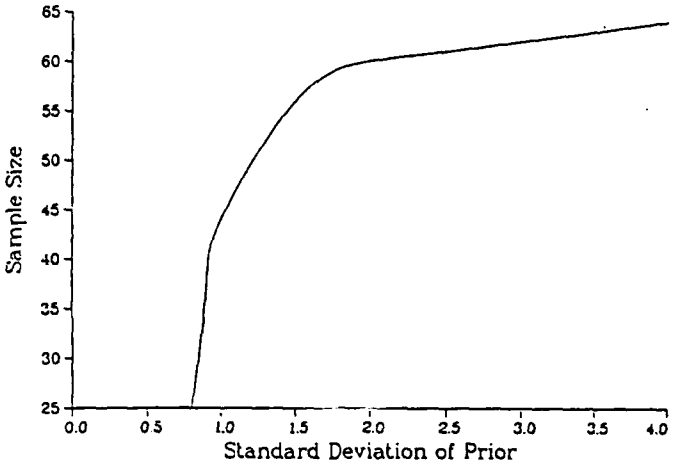


Figure 3-1. Plot of test sample size vs. prior standard deviation for Bayes modified minimax risk example.

4. TESTABILITY DEMONSTRATION

There are several possible measures of testability such as fraction of faults detected, false alarm rate, cannot duplicate rate, percent fault coverage and fraction faults isolatable (Klion, 1985). The variable fraction of faults isolatable (FFI) is used in this report, however the method is applicable to any measure for which detection can be measured as a Bernoulli random variable. The remainder of this report is restricted to the testability application. The following notation and assumptions are used.

4.1 Notation

n	sample size
Y	number of faults isolated in a test sequence
π	true FFI of the BIT design
$h(\pi)$	prior density function of
$g(Y \mid \pi)$	conditional density function of Y given
π_0	contract specified fault coverage
π_1	highest fault coverage for which $\pi \leq \pi_1$ results in total loss to the consumer
H_0:	the statistical hypothesis to be tested
$\delta(Y)$	Bayes decision function for a binomial test, i.e., $\delta(Y) = \begin{cases} a_0 \text{ if } Y > K \text{ (accept } H_0) \\ a_1 \text{ if } Y \leq K \text{ (reject } H_0) \end{cases}$
K	test critical value
$L(\pi, \delta(Y))$	Loss function for decision $\delta(Y)$ when FFI = π
$L_C(\pi)$	$L(\pi, \delta(Y) \mid \delta(Y)=a_0)$ (consumer loss function)
$L_P(\pi)$	$L(\pi, \delta(Y) \mid \delta(Y)=a_1)$ (producer loss function)
$E_\Delta(\cdot)$	expected value with respect to random variable Δ

4.2 Assumptions

Given a system with N possible states or responses each of which may be faulty, it is assumed that the number of potential faults (N) in the system is very large. The number (n) of faults simulated or observed in a test is assumed to be a small fraction

of the total possible in the system (N). It is further assumed that the simulated or observed faults represent a random sample of the population of potential faults. Under these assumptions isolation of a fault is approximately a Bernoulli random variable. Thus the FFI (Y) is a binomial random variable. Since Y is a sufficient statistic for π, it will be used in the test developed in this report.

It should be noted that the definition of π_1 is somewhat different from that used in the non-Bayeian testability demonstrations. For the classical test, π_1 represents fault coverage which would result in "serious" loss as opposed to "total" loss for the Bayes test. Thus it seems that π_1 for the Bayesian test may be somewhat lower. The relationship between seriousness of loss and value of π is controlled by the shape of the loss function which is considered in section 5.

The consumer and producer in testability demonstrations have very different perceptions of risk. Therefore it is necessary to use the Bayes modified minimax risk form given in (3-4). For this case the consumer risk is

$$r_c(\delta) = L_c(\pi_0^*) P(Y > K | \pi_0^*) P(\pi < \pi_0) \qquad (4\text{-}1)$$

and the producer Bayes modified minimax risk is

$$r_p(\delta) = L_p(\pi_1^*) P(Y \leq K | \pi_1^*) P(\pi \geq \pi_0). \qquad (4\text{-}2)$$

where π_0^* and π_1^* are defined (as in 3-4) such that

$$\begin{aligned} &\underset{\pi < \pi_0}{h(\pi) L_c(\pi) P(Y \geq K | \pi)} + \underset{\pi \geq \pi_0}{h(\pi) L_p(\pi) P(Y < K | \pi)} \\ &\leq \underset{\pi_0^* < \pi_0}{h(\pi_0^*) L_c(\pi_0^*) P(Y \geq K | \pi_0^*)} + \underset{\pi_1^* \geq \pi_0}{h(\pi_1^*) L_p(\pi_1^*) P(Y \geq K | \pi_1^*)} \end{aligned} \qquad (4\text{-}3)$$

5. LOSS FUNCTION

Ideally the loss function is an exact mathematical model of the economic (or other) penalties the consumer and producer experience when decision errors are made. In practice it is rarely possible to achieve this ideal. However it is possible to define a very general form which is much more realistic than the 0-1 loss function presently being used. The 0-1 loss function ignores losses in the region $\pi_1 < \pi < \pi_0$. In most applications this is a very likely region for the testability measure to occupy. Thus losses in this region should be important in experimental design. It seems inappropriate to use risk management strategies which ignore these losses.

The purpose of this section is to define the general character of the loss function as it relates to consumer and producer. A few forms are presented which are consistent with the loss information usually available.

In the context of testability, there are two opposing interests, the consumer and the producer. The assumption is made here that losses of the consumer dominate when a decision error results in acceptance of a system with $\pi < \pi_0$. Similarly, producer loss dominates whenever $\pi \geq \pi_0$ and the system is rejected. Therefore it is only necessary to consider consumer loss for $\pi < \pi_0$, and producer loss when $\pi \geq \pi_0$. This does not mean that the producer does not experience a loss if a bad system is accepted. For example, the producer may loose future contracts because of the resulting bad reputation. However because the consumer loss dominates, adequate protection against both losses is provided when the consumer loss alone is considered. This assumption permits representation of loss as

$$L(\pi, \delta(Y)) = \begin{cases} L_c(\pi) & \text{if } \delta(Y) = a_0 \\ L_p(\pi) & \text{if } \delta(Y) = a_1 \end{cases} \qquad (5\text{-}1)$$

where the consumer loss $L_c(\pi)$ is 0 for $\pi \geq \pi_0$ and the producer loss $L_p(\pi)$ is 0 if $\pi < \pi_0$.

A general function which applies to all tests must avoid unique reference monetary loss. The loss functions given in this report represent loss as a proportion of total system value. It is assumed that π_1 is defined so that the system value to the consumer is lost if $\delta(Y) = a_1$, i.e., $L_c(\pi_1) = 1$. Similarly if the system is rejected and $\pi \geq \pi_0$, $L_p(\pi) = 1$. If no decision error is made the loss to both consumer and producer is assumed to be 0.

For $\pi_1 < \pi < \pi_0$ the consumer loss is monotone decreasing. Two general forms are provided in this report. The first is given by the function

$$L'_c(\pi) = \begin{cases} 0 & \pi \geq \pi_0 \\ ((\pi - \pi_1)/(\pi_0 - \pi_1))^p & \pi_1 < \pi < \pi_0 \\ 1 & \pi \leq \pi_1 \end{cases} \qquad (5\text{-}2)$$

where the power p is to be specified. It would seem that a few values such as p = 1,2,3 could provide adequate choice in most applications. These functional shapes are illustrated in figure 5-1.

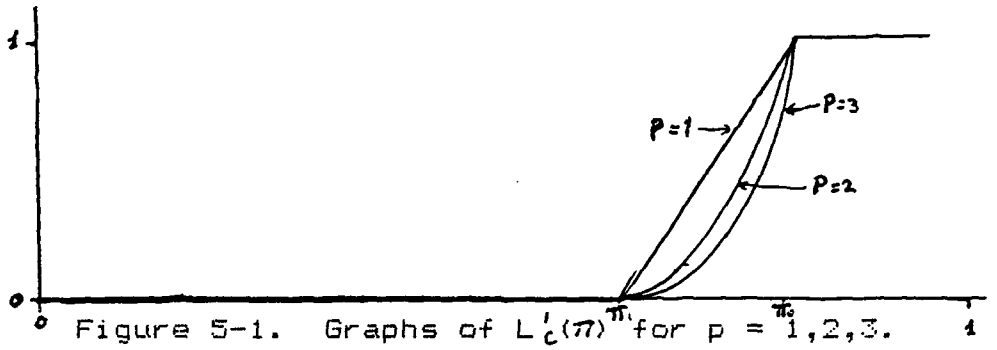


Figure 5-1. Graphs of $L_C^1(\pi)$ for p = 1,2,3.

The second general form is

$$L_C^2(\pi) = \begin{cases} 0 & \pi \geq \pi_0 \\ (p+1)\left((\pi_0 - \pi)/(\pi_0 - \pi_1)\right)^p - p\left((\pi_0 - \pi)/(\pi_0 - \pi_1)\right)^{p+1} & \pi_1 < \pi < \pi_0 \\ 1 & \pi \leq \pi_1 \end{cases} \quad (5-3)$$

where again p is to be specified. The form of $L_C^2(\pi)$ is illustrated in figure 5-2 for p = 2. It should be necessary to provide design tables for only a few values of p for both (5-2) and (5-3).

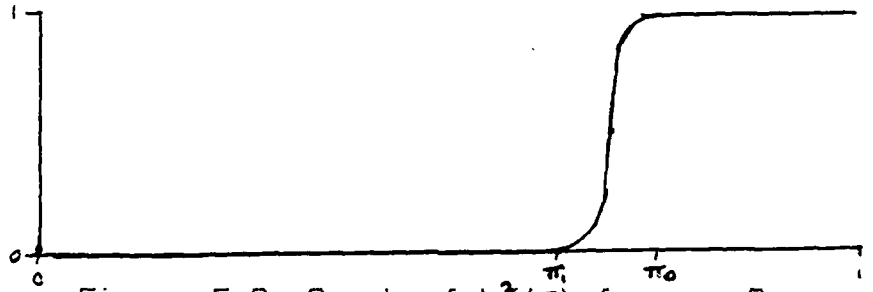


Figure 5-2. Graph of $L_C^2(\pi)$ for p = 2.

Because the consumer loss dominates when $\pi < \pi_0$, it is only necessary to define the producer loss for $\pi \geq \pi_0$. Rejection of the system in this case means total loss. Thus the producer loss function is the simple 0-1 form

$$L_P(\pi) = \begin{cases} 0 & \delta(Y) = a_0 \\ 1 & \delta(Y) = a_1 \end{cases}. \quad (5-5)$$

6. THE PRIOR DISTRIBUTION

The prior density $h(\pi)$ represents a quantification of prior information on the parameter π. Choosing the prior is a major stumbling block in many applications of Bayesian methods. Most procedures start with a "rich" family of distributions which represents all possible prior information configurations. The beta distribution family is such a family which also has a convenient mathematical relationship with the binomial distribution of FFI. This generality is very often one of the complicating factors in choosing $h(\pi)$.

In most applications prior information is not well defined and therefore does not require such a general family. An intuitively appealing method of selecting the prior given by Berger (1985), is to subjectively (or otherwise) estimate several fractiles of the prior density, i.e., points $\pi^*(\alpha)$ such that

$$P(\pi \leq \pi^*(\alpha)) = \alpha .$$

The prior is determined by fitting the functional form as close as possible to these fractiles. This method can be simplified considerably when used with the Bayes modified minimax risk criterion. This criterion permits the use of very simple prior distributions which address in a straight forward and intuitive manner the difficult problem of quantification of prior information.

There are two primary benefits of this approach. First, as already mentioned, it is very simple. Second, it permits the use of a few tables to determine the experimental design in any test situation. This approach should be compared with methods which require application of a computer program, or a complete volume of tables.

6.1 The Prior Density in Bayes Modified Minimax Risk

Before considering the form of the prior density, it is instructive to examine the relationship between the prior and Bayes modified minimax. Note from (4-2) and (4-3) that shape of $h(\pi)$ is not a consideration (except in the determination of π_0^* and of π_1^*). Thus the only aspect of the prior distribution which the producer and consumer need agree upon is the division of total probability between consumer and producer. The shape is not a consideration at this point.

Let $P_0 = P(\pi < \pi_0)$ and $P_1 = P(\pi \geq \pi_0)$. Given agreement on these values the producer and consumer are free to choose a shape which reflect their unique risk concerns.

6.1.1 Producer prior. Consider the producer risk. Given imprecise prior information it is not reasonable for the producer to choose a prior density with a pronounced peak in the region $\pi \geq \pi_0$. A producer conservative prior places more probability in the neighborhood of π_0 where error protection from the sample (represented by $P(Y < K | \pi)$ is smallest in the region $\pi \geq \pi_0$. Note that any prior which is monotone decreasing for $\pi \geq \pi_0$ will have

$$\max_{\pi \geq \pi_0} h(\pi) L_p(\pi) P(Y \leq K | \pi) = h(\pi_0) P(Y \leq K | \pi_0)$$

for the loss function (5-5). Thus for any producer conservative prior (i.e., monotone decreasing for $\pi \geq \pi_0$) $\pi_1^* = \pi_0$. Thus equation (4-2) may be written

$$r_p(\delta) = P(Y \leq K | \pi_0) P_1 \qquad (6\text{-}1)$$

6.1.2 Consumer prior. The consumer case is slightly more complicated. There are two regions of π which require individual attention, i.e., $\pi \leq \pi_1$ and $\pi_1 < \pi < \pi_0$. Let

$$M(\pi) = h(\pi) L_c(\pi) P(Y > K | \pi). \qquad (6\text{-}2)$$

Note that $P(Y > K|\pi)$ is a monotone increasing function of π. Define π_{01}^* and π_{02}^* such that

$$M(\pi_{01}^*) = \max_{\pi \leq \pi_1} M(\pi) \qquad (6\text{-}3)$$

and

$$M(\pi_{02}^*) = \max_{\pi_1 < \pi < \pi_0} M(\pi). \qquad (6\text{-}4)$$

Then in (4-1) (consumer risk)

$$\pi_0^* = \begin{cases} \pi_{01}^* & \text{if } M(\pi_{01}^*) \geq M(\pi_{02}^*) \\ \pi_{02}^* & \text{if } M(\pi_{01}^*) < M(\pi_{02}^*) \end{cases}$$

The Bayes minimax risk criteria simplifies choice of π_0^*. In most applications (having unimodal prior density) the prior density will be monotone increasing in a neighborhood of π_1. It follows since $L(\pi)P(Y > K|\pi)$ is a continuous function of π that

$$M(\pi_{01}^*) \leq M(\pi_{02}^*).$$

The solution for π_{02}^* in (6-4) requires numerical methods which depend upon the shape of $h(\pi)$ in the region $\pi_1 < \pi < \pi_0$. A uniform prior is used here for $\pi_1 < \pi < \pi_0$. It is suggested because it seems consistent with the quality of prior information commonly encountered. However it is a simple task to break the interval into a few segments and permit the analyst free choice of density values for a piecewise uniform density for $\pi_1 < \pi < \pi_0$.

It is important to emphasize that the previous discussion has dealt only with shape. The prior probabilities $P(\theta < \theta_0)$ and $P(\theta \geq \theta_0)$ have not been fixed. This simplification permits the use of a few tables representing a wide range of prior information states.

7. TESTABILITY DEMONSTRATION FORMAT

The Bayes modified minimax risk criterion is applied to the testability problem in this section. The loss functions and prior densities developed in the previous sections are used.

Given the binomial distribution of Y (number of faults isolated) the risks (4-1) and (4-2) become

$$r_c = P_0 L_c(\pi_0^*) P(Y > K | \pi_0^*) \qquad (7\text{-}1)$$

and

$$r_p = P_1 L_p(\pi_1^*) P(Y \leq K | \pi_1^*) \qquad (7\text{-}2)$$

The constant K represents the critical value such that the hypothesis is rejected if the observed $Y \leq K$; otherwise the hypothesis is accepted. After choosing the the loss function and prior density, the design problem is to select K and n so that the consumer and producer are satisfied with their respective risks r_c and r_p. Tables used to determine K and n are presented in this section. The tables provided represent only a sample of those which may be necessary for a complete testability demonstration program.

7.1 Tables for Design of Testability Demonstration

It can be noted from section 4 that the Bayes modified minimax risk depends upon π_0, π_1, P_0, P_1 and the loss function. The following transformations permit the use of tables which are not specific to P_0 and P_1 and thus independent of the prior density. Let

$$ar_c = r_c / P_0 \quad \text{and} \quad ar_p = r_p / P_1 \qquad (7\text{-}3)$$

represent "adjusted" risks for consumer and producer respectively.

The FORTRAN program BMMRSK.FOR listed in Appendix B computes design tables for testability demonstration. The program permits specification of loss function from among those described in section 5. Additional inputs required to generate a table are π_0 and π_1. A uniform prior for $\pi_1 < \pi < \pi_0$ is used. However it is a simple change of the computer code to permit a piecewise uniform prior density in the interval $\pi_1 < \pi < \pi_4$. However this generality greatly magnifies the number of required tables.

The program determines ar_c and ar_p for a range of n and K values. The following examples illustrate use of the tables for designing a testability demonstration.

7.1.1 Testability demonstration design example 1. Suppose contract specifies $\pi_0 = 0.90$ and it is determined that the system is essentially unusable if $\pi \leq 0.80$. The loss functions (5-2) with p = 3 and (5-5) are to be used. The producer has successfully designed and constructed similar testability equipment under previous contracts. Field records indicate the specified testability was achieved for these systems. It is subjectively

determined that the prior probability of the producer achieving the specified testability (P_1) is 0.75. Producer and consumer independently determine their respective bounds on allowable risk to be r_p = 0.10 and r_c = 0.05. Therefore adjusted risks become ar_p = 0.133 and ar_c = 0.20. Table A-1 in Appendix A contains values of ar and ar for various test designs (i.e., K and n). The values of K and n which are appropriate for this example are the smallest K and n for which both $ar_p \leq 0.133$ and $ar_c \leq 0.20$. For this case table A-1 indicates K = 49 and n = 58.

7.1.2 Testability demonstration design example 3. Consider a situation as in example 1 except that there has been very little experience with the producer. The least informative prior in this case is $P_0 = P_1 = 0.50$. The adjusted risks are ar_p = 0.20 and ar_c = 0.10. Then from table A-1, K=53 and n = 62.

7.1.3 Testability demonstration design example 3. Let π_0 = 0.90 and π_1 = 0.70 for this example. The risk bound 0.05 for both producer and consumer is to be used. Assume that the consumer has had successful association with producer. However because the producer is entering a new application area, It is felt that P_1 should be no higher than 0.60. The adjusted risks for this case become ar_c = 0.125 and ar_p = 0.083. Table A-2 yields the design K = 22 and n = 28.

listed in the following tables.

Table 8.1. Sensitivity analysis for section 7.1.1 example.
(π_0 = .90, π_1 = .80, r_c = .05, r_p = .10, P_0 = .25, P_1 = .75)
(Design sample size n = 58, critical value k = 49)

	Monte Carlo Estimates (1000 Samples)		
Prior Mode	P_0	r_c	r_p
.75	.999	.058	.000
.77	.997	.074	.000
.79	.994	.078	.000
.81	.975	.093	.001
.83	.918	.064	.009
.85	.855	.062	.005
.87	.755	.043	.010
.88	.699	.033	.011
.89	.614	.019	.020
.90	.515	.010	.019
.91	.399	.005	.029
.92	.224	.001	.031
.93	.084	.000	.030
.94	.015	.000	.020

Table 8.2. Sensitivity analysis for section 7.1.2 example.
(π_0 = .90, π_1 = .80, r_c = .05, r_p = .10, P_0 = P_1 = .50)
(Design sample size n = 58, critical value k = 49)

	Monte Carlo Estimates (1000 Samples)		
Prior Mode	P_0	r_c	r_p
.75	.999	.027	.000
.77	.999	.043	.000
.79	.994	.044	.000
.81	.975	.056	.001
.83	.935	.043	.006
.85	.844	.042	.009
.87	.740	.026	.021
.88	.678	.025	.022
.89	.593	.017	.015
.90	.493	.011	.031
.91	.382	.003	.031
.92	.212	.001	.044
.93	.089	.000	.031
.94	.021	.000	.022

8. TEST SENSITIVITY AND COMPARISON WITH NONBAYESIAN DESIGN

The sensitivity of the Bayes modified minimax approach to the chosen prior distribution is investigated in this section using Monte Carlo techniques. The effect of using prior information is also evaluated by comparing the test designs of Bayes and classical methods.

8.1 Sensitivity Analysis

Traditional sensitivity analysis investigates the operational characteristics of the test when improper priors are used. Such analysis is almost an academic exercise with the Bayes modified minimax risk criterion. It is the nature of minimax methods to protect against the worst case. Therefore any situation cannot be worse than that which was anticipated in the test. The Bayes modified minimax approach anticipates the worst case according to the likelihoods suggested by the prior information. Thus some sensitivity can be expected.

It is useful to review motivation for the Bayes modified approach to test design in order to develop a meaningful sensitivity analysis. Recall that the prior density is seldom the true density of π in any application. It seems reasonable that in most testability applications the contract specified value represents a target. The precise value of π achieved by the design is very likely unknown to the producer. In this sense it can be regarded as a true random variable. However in any real application, this random variable will have a very small variance compared with the conservative prior that is actually used to design the test. For this reason the prior is more appropriately regarded as a weighting useful for modifying the "absolute" worst case as defined by minimax methods. It is more informative to investigate sensitivity with respect to the type of distributions which are likely to be encountered in a typical application.

A truncated normal is used here to represent the prior density of π. The natural bounds, $0 \le \pi \le 1$ are used to truncate the normal distribution. It is also more realistic to consider a relationship between the mode and variance of the prior. For priors with very high mode, e.g., 0.95, it seems likely that the density will drop off more rapidly than cases for which the mode is relatively small. Otherwise the truncated normal will have high density values even at $\pi = 1.0$ which is unlikely in most applications. In fact when a testability value such as π_0 is a design goal it very unlikely that the true testability for the system will be much higher than π_0. This behavior of the prior can be modeled by reducing the prior variance as the mode of the becomes large. In the sensitivity analyses of this section a prior with mode m has the standard deviation (before truncation) which is the minimum of .05 and $(1 - m)/3$.

The program MONTE listed in appendix C is a Monte Carlo evaluation of the consumer and producer risks for a given testability demonstration design. This program was used to evaluate the risk sensitivity for the test situations described in the examples of sections 7.1.1,7.1.2, and 7.1.3. The results are

Table 8.3. Sensitivity analysis for section 7.1.3 example.
(π_c = .90, π_i = .70, r_c = r_p = .05, P_0 =.40, P_1 = .60)
(Design sample size n = 28, critical value k = 22)

	Monte Carlo Estimates (1000 Samples)		
Prior Mode	P_o	r_c	r_p
.75	.999	.093	.001
.77	.996	.085	.000
.79	.995	.074	.000
.81	.974	.052	.001
.83	.936	.039	.002
.85	.856	.026	.004
.87	.791	.013	.004
.88	.693	.009	.012
.89	.609	.005	.007
.90	.530	.002	.010
.91	.349	.001	.013
.92	.223	.000	.019
.93	.089	.000	.008
.94	.014	.000	.006

It is evident from the results of the Monte Carlo studies shown in the above tables that within a very wide range of P values, the estimated risks are well within the bounds imposed for the experimental design. It is also evident that the producer risk is very conservative. This result seemed sufficiently conservative that a check on the programs was made by executing the Monte Carlo program with the prior mode set at the Bayes modified "worst case" for both consumer and producer. The prior variance was set at 0. The resulting Monte Carlo estimated consumer and producer risks were very consistent with their respective values in the appropriate design tables. This consistent result supports the accuracy of this sensitivity analysis. It also suggests that the true risks are sensitive to the prior variance, the larger the prior variance, the more conservative the design. This conclusion is consistent with the conclusions of section 2, in which it was noted that a less informative prior (i.e., large variance) requires a smaller sample size when using existing design criteria. Since the prior distribution used for design purposes is in no way representative of the distribution of π (even if it can be considered a random variable), the conservatism imposed by the Bayes modified minimax approach seems justified.

8.2 Comparison with NonBayesian Designs

In order to compare the Bayesian and classical designs it is necessary to examine risk criteria for both cases. The classical approach can be characterized as a Bayes design in which two "degenerate" priors are used. For the consumer case the prior loads all probability at π_i. (There is some difference in interpretation of π_i between the Bayesian and classical tests as noted in section 4.) For the producer case the classical test loads

all probability at π_0. Thus the adjusted risks (7-3) of the Bayes approach are comparable with the classical risks when the prior density is such that $P_0 = P_1 = 0.50$. There is not a perfect comparison because the Bayes approach uses a loss function. However for the loss function used in examples 7.1.1 and 7.1.2 the minimax solutions do not deviate a great deal from the values π_0 and π_1.

Consider a classical fixed sample test of the hypothesis

$$H_0 : \pi \geq .90 \text{ vs. } H_1 : \pi < .90$$

Let $\alpha(\pi_0) = 0.20$ and $\beta(\pi_1) = 0.10$ be used for the type I and type II error bounds (which are the respective adjusted risks for example 7.1.2.) The resulting design is n = 62 with K = 53 which is the same as the design for example 7.1.2.

The design for example 7.1.1 can now be evaluated for the effect of the prior information. The design for this example is n = 55 with K = 46. Use of the prior information resulted in an 11% reduction in required sample size.

9. CONCLUSIONS AND RECOMMENDATIONS

Existing Bayes methods for experimental design are inconsistent with respect to influence of the prior information. It has been shown that a less informative prior can produce a design with smaller sample size than that of a more informative prior. This result would be valid if the parameter were actually a random variable with a distribution which can be estimated. In reality the prior is an expression of ignorance with respect to the parameter and requires a conservative approach. The notion of conservatism in Bayesian estimation does not carry over to experimental design.

The Bayes modified minimax risk criterion is introduced to provide a consistent and more realistic approach to experimental design. This approach greatly simplifies the task of quantifying prior information. Tables have been prepared to illustrate the design method. Every effort has been made here to simplify the inputs required of the analyst. In particular a very simple form for the prior density is used. The mathematical structure of the Bayes modified approach permits easy extension to more general prior densities. The extension will of course require greater sophistication of the analyst and more extensive tables. The intent here is to reduce the complexity of Bayes methods and hence encourage their use.

The illustrations used in this report serve as examples only. It is recommended that consideration be given to the loss functions and prior densities which are adequate for testability demonstration. If it is determined that more general forms should be made available, then the program for generating tables should be modified to include these forms. For example it is not necessary that the loss function and prior density share the boundaries (e.g., π_0 and π_1) for definition of the function. Also rather than a uniform shape for $\pi_1 < \pi < \pi_0$, it is possible to permit 2 or 3 levels (1 or 2 steps) in the density function without a great deal of added complexity.

The problem with existing Bayesian experimental design methods revealed in this report extends to applications other than testability demonstration. Therefore it is recommended that other Bayesian design applications be reviewed to the advisability of using the Bayes modified minimax approach.

10. REFERENCES

Berger, J.O., 1985. Statistical Estimation Theory and Bayesian Analysis. Springer-Verlag, New York.

Canfield, R.V. and Teed, J.C., 1977. Selecting the prior distribution in Bayesian statistics. IEEE Trans. on Reliability, R-26:283-285.

Easterling, R.G., 1970. On the use of prior distributions in in acceptance sampling. Annals of Reliability and Maintainability, Vol. 9:31-35.

Goel, A.L. and Joglekar, A.M., 1976. Reliabiltiy Acceptance Sampling Plans Based Upon Prior Distribution, Risk Criteria and Their Interpretation. Griffiss AFB, NY, Rome Air Development Center, RADC-TR-76-294, Vol. 2.

Goel, A.L. and Joglekar, A.M., 1976. Reliability Acceptance Sampling Plans Based Upon Prior Distribution, Sensitivity Analysis. Griffiss AFB, NY, Rome Air Development Center, RADC-TR-76-294, Vol. 5.

Heiser, D.A., 1986. New techniques in R&M contract requirements. Proceedings of the Annual Reliability and Maintainabiltiy Symposium.

Jaynes, E.T., 1968. Prior probabilities. IEEE Trans. on System Science and Cybernetics, SSC-4:227-241.

Klion, J., 1985. A Rational and Approach for Defining and Structuring Testability Requirements, Griffiss AFB, Rome, NY, Rome Air Development Center, RADC-TR-85-150.

Martz, H.F. and Waller, R.A., 1982. Bayesian Reliability Analysis. Wiley, New York.

Schick, G.J. and Drnas, T.M., 1972. Bayesian reliability demonstration, AIIE Transactions, 4:92-102.

APPENDIX A

Design Tables

Table A-1

BAYES MODIFIED MINIMAX RISK
PIE0 = 0.90 PIE1 = 0.80
LOSS FUNCTION TYPE = 1 POWER = 3.0

			K+ 0	K+ 1	K+ 2	K+ 3	K+ 4	K+ 5	K+ 6	K+ 7	K+ 8
N	50	APR	.0094	.0245	.0579	.1221	.2298	.3839	.5688	.7497	
K	39	ACR	.5836	.4437	.3073	.1904	.1034	.0491	.0204	.0071	
N	51	APR	.0038	.0109	.0279	.0643	.1329	.2452	.4024	.5869	.7636
K	39	ACR	.6852	.5556	.4165	.2839	.1730	.0924	.0436	.0180	.0062
N	52	APR	.0045	.0126	.0315	.0712	.1441	.2609	.4208	.6046	.7768
K	40	ACR	.6593	.5278	.3900	.2618	.1569	.0825	.0387	.0159	.0055
N	53	APR	.0053	.0145	.0355	.0785	.1558	.2769	.4392	.6218	.7895
K	41	ACR	.6330	.5002	.3643	.2408	.1420	.0738	.0344	.0140	.0048
N	54	APR	.0063	.0166	.0398	.0862	.1679	.2931	.4575	.6386	.8015
K	42	ACR	.6064	.4730	.3396	.2210	.1282	.0661	.0306	.0124	.0042
N	55	APR	.0073	.0189	.0444	.0944	.1804	.3096	.4756	.6549	.8130
K	43	ACR	.5798	.4463	.3159	.2025	.1156	.0592	.0273	.0110	.0037
N	56	APR	.0084	.0214	.0494	.1030	.1934	.3262	.4935	.6707	
K	44	ACR	.5531	.4203	.2932	.1851	.1041	.0530	.0243	.0097	
N	57	APR	.0036	.0097	.0242	.0548	.1120	.2066	.3429	.5112	.6860
K	44	ACR	.6527	.5265	.3948	.2716	.1689	.0939	.0476	.0217	.0086
N	58	APR	.0042	.0112	.0273	.0605	.1215	.2203	.3597	.5287	.7008
K	45	ACR	.6275	.5002	.3702	.2510	.1538	.0847	.0427	.0193	.0076
N	59	APR	.0049	.0128	.0306	.0666	.1314	.2342	.3766	.5459	.7152
K	46	ACR	.6020	.4742	.3464	.2316	.1398	.0765	.0383	.0172	.0067
N	60	APR	.0057	.0146	.0342	.0731	.1416	.2484	.3935	.5628	.7290
K	47	ACR	.5764	.4486	.3234	.2132	.1268	.0691	.0345	.0154	.0060
N	61	APR	.0066	.0165	.0381	.0799	.1523	.2630	.4105	.5795	.7424
K	48	ACR	.5509	.4236	.3014	.1959	.1152	.0624	.0310	.0138	.0053
N	62	APR	.0076	.0187	.0423	.0872	.1634	.2777	.4274	.5957	.7553
K	49	ACR	.5254	.3991	.2803	.1797	.1046	.0564	.0279	.0123	.0047
N	63	APR	.0087	.0211	.0468	.0948	.1748	.2927	.4442	.6117	.7676
K	50	ACR	.5002	.3754	.2602	.1645	.0951	.0511	.0251	.0110	.0042
N	64	APR	.0038	.0099	.0236	.0516	.1028	.1866	.3078	.4610	.6273
K	50	ACR	.5981	.4752	.3523	.2410	.1504	.0865	.0462	.0226	.0099
N	65	APR	.0044	.0113	.0264	.0567	.1112	.1987	.3231	.4776	.6425
K	51	ACR	.5735	.4506	.3301	.2229	.1374	.0787	.0419	.0203	.0088
N	66	APR	.0051	.0128	.0295	.0621	.1199	.2112	.3386	.4941	.6574
K	52	ACR	.5489	.4265	.3086	.2058	.1256	.0716	.0379	.0183	.0079
N	67	APR	.0059	.0145	.0327	.0679	.1290	.2239	.3541	.5104	.6718
K	53	ACR	.5244	.4029	.2881	.1896	.1148	.0652	.0344	.0165	.0071
N	68	APR	.0067	.0163	.0362	.0740	.1385	.2369	.3698	.5266	.6859
K	54	ACR	.5001	.3800	.2684	.1744	.1050	.0594	.0312	.0149	.0063
N	69	APR	.0077	.0183	.0400	.0805	.1484	.2502	.3854	.5425	.6995
K	55	ACR	.4761	.3576	.2496	.1602	.0961	.0541	.0283	.0134	.0057
N	70	APR	.0088	.0205	.0441	.0873	.1586	.2637	.4011	.5582	.7128
K	56	ACR	.4524	.3360	.2317	.1472	.0880	.0494	.0256	.0121	.0051
N	71	APR	.0099	.0228	.0484	.0944	.1691	.2775	.4169	.5736	.7256
K	57	ACR	.4291	.3152	.2147	.1353	.0806	.0450	.0233	.0109	.0046
N	72	APR	.0046	.0112	.0254	.0530	.1019	.1799	.2914	.4325	.5888
K	57	ACR	.5236	.4063	.2951	.1986	.1245	.0738	.0411	.0211	.0099
N	73	APR	.0053	.0126	.0281	.0579	.1097	.1911	.3055	.4482	.6038
K	58	ACR	.5001	.3841	.2758	.1835	.1145	.0677	.0375	.0192	.0089
N	74	APR	.0060	.0142	.0311	.0631	.1178	.2025	.3198	.4637	.6184
K	59	ACR	.4769	.3624	.2573	.1694	.1054	.0620	.0342	.0174	.0081
N	75	APR	.0068	.0159	.0343	.0685	.1263	.2142	.3342	.4792	.6327
K	60	ACR	.4540	.3414	.2397	.1565	.0970	.0569	.0312	.0158	.0073

Table A-2

BAYES MODIFIED MINIMAX RISK
PIE0 = 0.90 PIE1 = 0.70
LOSS FUNCTION TYPE = 1 POWER = 3.0

			K+ 0	K+ 1	K+ 2	K+ 3	K+ 4	K+ 5
N	25	APR	.0095	.0334	.0980	.2364	.4629	.7288
K	18	ACR	.3407	.1935	.0953	.0406	.0141	.0035
N	26	APR	.0119	.0399	.1118	.2591	.4895	.7487
K	19	ACR	.2965	.1637	.0801	.0339	.0116	.0029
N	27	APR	.0039	.0147	.0471	.1266	.2821	.5154
K	19	ACR	.4113	.2563	.1391	.0676	.0284	.0096
N	28	APR	.0050	.0179	.0550	.1421	.3054	.5406
K	20	ACR	.3648	.2202	.1186	.0573	.0238	.0080
N	29	APR	.0062	.0216	.0637	.1584	.3290	.5650
K	21	ACR	.3214	.1894	.1014	.0487	.0201	.0067
N	30	APR	.0078	.0258	.0732	.1755	.3526	.5886
K	22	ACR	.2814	.1634	.0870	.0415	.0170	.0056
N	31	APR	.0096	.0306	.0834	.1932	.3762	.6114
K	23	ACR	.2450	.1413	.0749	.0354	.0144	.0047
N	32	APR	.0117	.0358	.0944	.2115	.3997	.6333
K	24	ACR	.2134	.1226	.0646	.0304	.0122	.0040
N	33	APR	.0141	.0417	.1061	.2303	.4231	.6543
K	25	ACR	.1865	.1066	.0558	.0261	.0104	.0033
N	34	APR	.0051	.0169	.0481	.1185	.2496	.4462
K	25	ACR	.2679	.1633	.0929	.0484	.0225	.0089
N	35	APR	.0063	.0200	.0552	.1316	.2693	.4690
K	26	ACR	.2360	.1433	.0812	.0420	.0194	.0076
N	36	APR	.0077	.0235	.0628	.1454	.2892	.4915
K	27	ACR	.2084	.1260	.0710	.0366	.0168	.0065
N	37	APR	.0093	.0274	.0711	.1598	.3095	.5136
K	28	ACR	.1844	.1111	.0623	.0319	.0145	.0056
N	38	APR	.0111	.0318	.0800	.1747	.3299	.5352
K	29	ACR	.1635	.0981	.0547	.0279	.0126	.0048
N	39	APR	.0131	.0366	.0894	.1903	.3504	.5563
K	30	ACR	.1452	.0867	.0482	.0244	.0109	.0042
N	40	APR	.0051	.0155	.0419	.0995	.2063	.3710
K	30	ACR	.2046	.1291	.0768	.0425	.0214	.0095
N	41	APR	.0061	.0181	.0477	.1102	.2227	.3916
K	31	ACR	.1830	.1151	.0682	.0375	.0188	.0083
N	42	APR	.0073	.0211	.0539	.1214	.2396	.4121
K	32	ACR	.1639	.1027	.0606	.0332	.0165	.0073
N	43	APR	.0087	.0244	.0607	.1333	.2569	.4325
K	33	ACR	.1469	.0917	.0539	.0294	.0145	.0063
N	44	APR	.0103	.0280	.0679	.1456	.2744	.4528
K	34	ACR	.1320	.0821	.0480	.0260	.0128	.0055
N	45	APR	.0120	.0320	.0757	.1585	.2923	.4729
K	35	ACR	.1187	.0735	.0428	.0231	.0113	.0049
N	46	APR	.0048	.0140	.0364	.0840	.1719	.3103
K	35	ACR	.1644	.1068	.0660	.0383	.0205	.0100
N	47	APR	.0058	.0163	.0411	.0928	.1857	.3286
K	36	ACR	.1486	.0963	.0592	.0342	.0183	.0088
N	48	APR	.0068	.0187	.0463	.1021	.2000	.3469
K	37	ACR	.1346	.0869	.0533	.0306	.0163	.0078
N	49	APR	.0080	.0215	.0519	.1119	.2147	.3654
K	38	ACR	.1219	.0785	.0479	.0274	.0145	.0069
N	50	APR	.0094	.0245	.0579	.1221	.2298	.3839
K	39	ACR	.1106	.0710	.0432	.0246	.0129	.0061

APPENDIX B

Listing of the FORTRAN program BMMRSK.FOR and subroutines which are used to generate design tables for testability demonstration.

```
C*****This program drives subroutines which compute tables  for
C*****determining sample size and critical value for the test
C*****Ho: pie >= pie0 of the binomial parameter pie.  Output
C*****is a table of producer risk/consumer risk at various
C*****sample sizes and critical values for the general case and
C*****producer risk/max(pie0-pie)*P(X <= crit.val.) for the minimax
C*****case.  Hypothisized probability of success is pie0 and the
C*****minimum acceptable probability of success is pie1.
        COMMON CC(6),C(2,2,2),P(2,2,2),CL(4,6),PL(4,6)
        COMMON ICLF,IPLF,ICP,IPP,H1,H2,POWER
        DIMENSION AC(20),AP(20),PM(20)
        CHARACTER FIRST*4/'+___'/,SEC*6/'______'/,THIRD*4/' ___'/
        CHARACTER BL1*10/'          '/,BL2*4/'  K+'/
        CHARACTER CN*2/' N'/,APR*4/'¦APR'/,CCC*2/' K'/
        CHARACTER PIE*4/'¦PIE'/,ACR*4/'¦ACR'/
        INTEGER STEP
C*****This subroutine is used to choose the desired loss function
C*****and prior distributions.
C*****NL and NU are lower and upper bounds of sample size for the
C*****table.  AL,BL are linear coefficients for determining
C*****lower  bounds for critical values to be computed
C*****at each sample size (N).
        PIE1=1.0
        CALL SETUP(RC,RP,PIE0,PIE1,AL,BL,NL,NU,STEP,ACRMIN)
        ICL=AL+BL*NU
        DO 50 I=ICL,NU-1
        CALL PROBS(NU,I,PIE0,PIE1,PR,CR,PIE)
        IF (CR.LT.ACRMIN) GO TO 51
 50     CONTINUE
 51     NDMAX=I-ICL+1
        IF(NDMAX.GT.20) NDMAX=20
        WRITE(16,104) BL1,((BL2,J),J=0,NDMAX-1)
 104    FORMAT(A10,16(A4,I2))
        WRITE(16,101) FIRST,(SEC,J=0,NDMAX)
        DO 2 N=NL,NU,STEP
        ICL=AL+BL*N
        ICU=ICL+19
        IF (ICU.GT.N-1) ICU=N-1
        ND=ICU-ICL+1
        CR=1.0
        DO 3 I=1,ND
        IF(CR.LT.ACRMIN) GO TO 5
        IC=ICL+I-1
C*****This subroutine computes the producer risk (PR) and
C*****consumer risk (CR) at sample size N and critical value
C*****IC for pie0 and pie1 for the general case.  It computes
C*****producer risk (PR) and max(pie0 - pie)*P(X <= IC) for
C*****the minimax case.
        CALL PROBS(N,IC,PIE0,PIE1,PR,CR,PMAX)
        PM(I)=PMAX
        AC(I)=CR
```

```
 3        AP(I)=PR
 5        ND=I-1
          JS=1
          IF (AP(2).LT.APMIN) JS=2
          IF (NDMAX-JS+1.LT.ND) JS=ND-NDMAX+1
          ICL=ICL+JS-1
          WRITE(16,100) CN,N,APR,(AP(J),J=JS,ND)
          WRITE(16,100) CCC,ICL,ACR,(AC(J),J=JS,ND)
C***      WRITE(16,102) SEC,PIE,(PM(J),J=JS,ND)
          WRITE(16,101) FIRST,(SEC,J=0,NDMAX)
 2        CONTINUE
 100      FORMAT(A2,I4,A4,20(X,F5.4))
 101      FORMAT(A4,21A6)
 102      FORMAT(A6,A4,20(X,F5.4))
          STOP
          END
```

```
C*****This subroutine computes the max producer risk (APR) and
C*****max consumer risk (ACR).
      SUBROUTINE PROBS(N,IC,PIE0,PIE1,APR,CMAX,PMAX)
      COMMON CC(5),C(2,2,2),P(2,2,2),CL(4,5),PL(4,5)
      COMMON ICLF,IPLF,ICP,IPP,H1,H2,POWER
      DIMENSION PM(5)
C*****This subroutine computes the cummulative binomial probability
C*****P(X <= IC) = PS.
      CALL BIN(IC,N,PIE0,PS,PIC)
      APR=PS
      CALL BIN(IC,N,PIE1,CS1,PIC)
      CC(1)=1.-CS1
      CMAX=CC(1)
      PMAX=PIE1
      PIE=PIE1
      PM(1)=PIE
      D=PIE0-PIE1
      DPIE=D/2.
C*****This section of the program is a search for
C*****max(Loss(pie))*P(X > IC).
 6    DPIE=DPIE*2./3.
      DO 1 I=1,2
      TPIE=PIE+DPIE*I
      DD=(PIE0-TPIE)/D
      IF (ICLF.EQ.1) DD=DD**POWER
      IF (ICLF.EQ.2) DD=(POWER+1.)*DD**POWER-POWER*DD**(POWER+1.)
      CALL BIN(IC,N,TPIE,PS,PIC)
      CC(I+1)=DD*(1.-PS)
      PM(I+1)=TPIE
      IF(CC(I+1).LT.CC(I)) GO TO 2
      CMAX=CC(I+1)
 1    PMAX=TPIE
      IF(DPIE.LE.0.00001) GO TO 4
      PIE=PIE+DPIE
      CC(1)=CC(2)
      PM(1)=PM(2)
      PMAX=PM(1)
      CMAX=CC(1)
 2    IF(DPIE.GT.0.00001) GO TO 6
 4    RETURN
      END
```

```
C*****This subroutine sets up loss functions and prior distributions
C*****for the MAXRISK case.  It also sets lower and upper bounds for
C*****sample size and critical value for computations.
      SUBROUTINE SETUP(RC,RP,PIE0,PIE1,AL,BL,NL,NU,STEP,ACRMIN)
      COMMON CC(5),C(2,2,2),P(2,2,2),CL(4,5),PL(4,5)
      COMMON ICLF,IPLF,ICP,IPP,H1,H2,POWER
      INTEGER STEP
      OPEN(UNIT=16,FILE='MAXRSK.DAT',STATUS='NEW')
      PRINT*,' ENTER MIN SAMPLE SIZE, MAX SAMPLE SIZE AND STEP'
      PRINT*,' SIZE FOR TABLE'
      READ(5,*) NL,NU,STEP
      print*,' enter minimum APR and minimum ACR for table'
      READ(5,*) APRMIN,ACRMIN
      PRINT*,' ENTER pie0,pie1'
      READ(5,*) PIE0,PIE1
      DO 10 IL=NL-1,1,-1
      CALL PROBS(NL,IL,PIE0,PIE1,PR,CR,PIE)
      IF(PR.LT.APRMIN) GO TO 11
10    CONTINUE
11    DO 12 IU=NU-1,1,-1
      CALL PROBS(NU,IU,PIE0,PIE1,PR,CR,PIE)
      IF(PR.LT.APRMIN) GO TO 13
12    CONTINUE
13    UN=NU-NL
      BL=(IU-IL)/UN
      AL=IU-BL*NU
      PRINT*,' ENTER TYPE OF CONSUMER LOSS FUNCTION FOR'
      PRINT*,' pie1 < pie < pie0'
      PRINT*,' 1 = (PIE0-PIE)**POWER'
      PRINT*,' 2 = S SHAPE'
      READ(5,*) ICLF
      PRINT*,' ENTER POWER FOR LOSS FUNCTION SHAPE'
      READ(5,*) POWER
      WRITE(16,102)
102   FORMAT(' BAYES MODIFIED MINIMAX RISK')
      WRITE(16,101) PIE0,PIE1
101   FORMAT(' PIE0 =',F5.2,' PIE1 =',F5.2)
30    WRITE(16,100) ICLF,POWER
100   FORMAT(' LOSS FUNCTION TYPE =',I2,' POWER =',F4.1)
      RETURN
      END
```

```
C*****This subroutine computes the cummulative binomial probability
C*****P(X <= KK) = PSS at sample size N and probability of success
C*****pie.
      SUBROUTINE BIN(KK,N,PIE,PSS,PK)
      REAL*16 PS,PI,PI1,P
      K=KK
      P=PIE
      IF (N*PIE.LE.KK) GO TO 3
      K=N-KK-1
      P=1.-PIE
 3    PS=(1.-P)**N
      PI=PS
      DO 1 I=0,K-1
      PI1=PI*P*(N-I)/((I+1.)*(1.-P))
      PS=PS+PI1
 1    PI=PI1
      PK=PI
      PSS=PS
      IF (N*PIE.GT.KK) PSS=1.-PSS
      RETURN
      END
```

APPENDIC C

Listing of the FORTRAN program MONTE.FOR which was used as a Monte Carlo evaluation of sensitivity in section 8.

```
      PRINT*,' ENTER RANDOM NUMBER SEED'
      READ(5,*) ISEED
10    PRINT*,' ENTER PIE0,PIE1'
      READ(5,*) PIE0,PIE1
      PRINT*,' ENTER POWER OF CONSUMER LOSS FUNCTION'
      READ(5,*) POWER
      PRINT*,' ENTER STAND. DEV. OF PRIOR'
      READ(5,*) SIG
      PRINT*,' ENTER N,K'
      READ(5,*) N,K
      PRINT*,' ENTER MC'
      READ(5,*) MC
      PRINT*,' ENTER LOWER AND UPPER PRIOR MODE, STEP'
      READ(5,*) SM,UM,STEP
      IF(UM.EQ.0.) STOP
      NN=(UM-SM)/STEP
      PRINT*,' PIE0 =',PIE0,' PIE1 =',PIE1,' N =',N,' K =',K
      PRINT*,' MODE      P0      RC      RP'
      DO 100 II = 1,NN
      AVE=SM+II*STEP
      STD=MIN(SIG,(1.-AVE)/3.)
      P0=0.
      P1=0.
      D=PIE0-PIE1
      C2=0.
      AN=0.
      RN=0.
      DO 1 I=1,MC
13    U = RAN(ISEED)
      U=ABS(U)

      T=SQRT(-2.*ALOG(U))
      X=T-(2.30753+.27061*T)/(1.+.99229*T+.04481*T*T)
      B=RAN(ISEED)
      IF(B.LE.0.50) X=-X
      X=STD*X+AVE
      IF(X.GT.1.0) GO TO 13
      IF(X.LT.PIE0) P0=P0+1.
      KK=0
      DO 2 J=1,N
      U=RAN(ISEED)
      IF(U.LE.X) KK=KK+1
2     CONTINUE
      IF(KK.LE.K) GO TO 11
      AN=AN+1.
      IF(X.GE.PIE0) GO TO 1
      COST=1.
      IF(X.GT.PIE1) COST=((PIE0-X)/D)**POWER
      C2=C2+COST
      GO TO 1
11    IF(X.GE.PIE0) P1=P1+1.
      RN=RN+1.
1     CONTINUE
      PR=P1/MC
      CR=C2/MC
      P0=P0/MC
      PRINT*,AVE,P0,CR,PR
100   CONTINUE
      GO TO 10
      END
```

Crystalline Silicon Electro-Optic Waveguides

Stephen R. Giguere

Prof. Lionel Friedman

Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA

23-January-1989

Richard A. Soref—Rome Air Development Center, Hanscom AFB, Massachusetts

---

Joseph P. Lorenzo—Rome Air Development Center, Hanscom AFB, Massachusetts

This project was sponsored by the Air Force Office for Scientific Research, Bolling Air Force Base, D.C.

Digital Equipment Corporation

CHAPTER 1

CRYSTALLINE SILICON ELECTO-OPTICAL WAVEGUIDES

Introduction

> Silicon, as an electronic substrate, sparked a technological revolution that allowed realization of very large scale integrated circuits. Silicon, as a mechanical substrate, promises to spark another technological revolution that will allow the realization of integrated *photonic* circuits. [1]

This work is a continuation of the research initiated by Prof. Lionel Friedman of Worcester Polytechnic Institue and Richard Soref and Joseph Lorenzo of the Rome Air Development Center at Hanscom Air Force Base in the field of silicon integrated optics. In their paper, Friedman, Soref and Lorenzo [2] proposed a number of novel longitudinal and transverse electro-optic modulators devices using silicon. Soref and Lorenzo have published articles on their work in silicon integrated optics, and many of the equations used in the mathematical modeling of the waveguides are based on their earlier findings on the optical properties of crystalline silicon. [3] [4]

Silicon as an Optical Media

Unlike fiber optic cables which propagate signals many kilometers, integrated optical waveguides need carry a signal only a few centimeters. Consequently, media that have been rejected in fiber optic applications may have practical applications in integrated optical circuitry. The most studied material in the field of integrated optics is probably $LiBNO_3$ [5] because it exhibits a strong linear electro-optic effect. [6] But silicon is also an attractive material for integrated optics research. Soref and Lorenzo have noted two advantages: [7]

1. Many of the processes developed for the Si electronics circuit industry can be applied to Si optical devices.

[1] Ronald Levy, "Integrated Photonic Devices on Silicon", Solid State Technology, November, 1988, p. 81.

[2] Lionel Friedman, Richard Soref and Joseph Lorenzo, "Silicon Double-injection Electro-optic Modulators with Junction Gate Control", Journal of Applied Physics, March 15, 1988, p. 1831-9.

[3] For example, see Richard Soref and Joseph Lorenzo, "All-Silicon Active and Passive Guided-Wave Components for 1.3 uM and 1.6 uM Wavelengths", IEEE Journal of Quantum Electronics, Vol QE-22, June 1986, p. 873.

[4] Richard Soref and Brian Bennett, "Electrooptical Properties in Silicon", IEEE Journal of Quantum Electronics, January, 1987, p. 123-8.

[5] For example, "$LiBNO_3$ High Speed Traveling Wave Electro-optic Waveguide Devices," A. R. Beaumont et al. or "High Sensitivity Band-pass rf Modulators in $LiBNO_3$," G. E. Betts, from the SPIE Conference in Boston, MA, vol. 993 September 7-9, 1988.

[6] Leon McCaughan, short course SPIE Conference, "Advanced Guided-wave Integrated Optic Devices," September 6, 1988, p. 3 class notes.

[7] Richard Soref and Joseph Lorenzo, "All-Silicon Active and Passive Guided-Wave Components for 1.3 uM and 1.6 uM Wavelengths", IEEE Journal of Quantum Electronics, p. 873

2. High speed Si electronic circuits can be combined monolithically with Si guided-wave devices in an optoelectronic integration.

Other researchers have also noted the optical properties of silicon as an optoelectronic media.

> Silicon is a unique substrate for planar and channel waveguide structures. It is widely available, inexpensive, chemically stable, and mechanically rugged. A sophisticated materials and process technology is already in existence. When polished or oxidized, its surface is ideal for film growth by deposition processes. Smooth topographic channel and ridge structures are readily formed by anisotropic etching. It is a good photodetector material and lends itself to monolithic integration with microelectronic circuits. [8]

Refractive Index Changes in Silicon

Unlike $LiBNO_3$, silicon does not exhibit a linear electro-optic (Pockel's) effect; therefore, smaller effects such as the Kerr effect, must be utilized in silicon optical switch design. Soref and Lorenzo have designed and built simple silicon electro-optic switches using the free carrier injection to induce a change in the index of refraction of the waveguide. [9]

To create an electro-optic modulator in silicon, one would need the ability to move electrons and holes around a confined waveguiding region to induce a carrier refraction effect. If a sufficient concentration of charge carriers can be injected or depleted, then the refractive index of the waveguide will be altered by Δn. An electro-optic modulator to phase modulate optical signals can be designed using index of refraction changes induced in the waveguide.

In design, "waveguide width, depth and refractive index (difference) between the waveguide and substrate are three parameters that control the waveguide properties (and) can be separately controlled in fabrication." [10] The physical dimensions of a waveguide cannot be changed but there are different effects, such as polarization, electro-optical and acousto-optical effects, which can contribute to a change in the index of refraction of a material. Therefore, the optical properties of a medium may be manipulated to design waveguide modulators and switches for optical communications. [11] Our goal was to model a novel silicon waveguiding structure where a sufficient change in the electron and hole concentrations in the waveguiding region could be modulated by an external applied voltage. Building on earlier work from Soref and Bennett, large electron (ΔN_d) and hole (ΔN_a) changes induce a refractive index change of Δn = 0.001, sufficient to create a Mach-Zehander Interferometer.

Numerical Methods

The earlier study on evaluating silicon integrated optic devices was made using a one dimensional model, meaning that certain assumptions were made to simplify the analysis: namely, waveguide models were analyzed assuming one-dimensional planar-current flow, diffusion currents were neglected and simplified treatment of recombination effects was employed. [12]

---

[8] Fred Hickernell, "Optical Waveguides on Silicon", Solid State Technology, November, 1988, p. 83.

[9] J. P. Lorenzo and Richard Soret, " 1.3 uM Electro-optic Silicon Switch", Applied Physics Letters, vol 51, July 6, 1987, p 6.

[10] R. C. Alferness, "Optical Guided-wave Devices", Science, November 14, 1986, p. 825.

[11] Amnon Yariv and Pochi Yeh, Optical Waves in Crystals, (Wiley and Sons, New York, 1984), p. 220.

[12] Lionel Friedman, "Proposal To Air Force Office of Scientific Research", Research Initiation Program, December, 1987, p. 1.

The goal of this research is to design and analyze a more complete model of silicon waveguide structures. The silicon waveguide model was designed and characterized using PISCES IIB, a two dimensional semiconductor device simulation program developed at Stanford University. PISCES IIB solves Poisson's equation and the continuity equations for electrons and holes to obtain the voltage potential, electron, hole and current density distributions for a given device geometry and set of doping and biasing conditions.

Using the electron and hole concentration data from the PISCES IIB simulation solutions, the predicted optical properties of the device were calculated. The refractive index changes at 1.3 uM and 1.55 uM wavelengths were calculated from data provided by Richard Soref, [13] and the index change information was used with a two dimensional representation of the cosine overlap integral to calculate the effective index change over the waveguiding region of a semiconductor.

A simple two terminal MOS diode was used to test the accuracy of the PISCES IIB solutions. Since single and double injection MOSFETs are also insulated gate structures, this simple case provides a strong foundation for analyzing more complex geometries.

Three terminal metal-oxide-semiconductor field effect transistors (MOSFETs) and double injection field effect transistors (DIFETs) were investigated to see if the free carrier concentrations could be manipulated in a useful manner. Maximum current densities within the device were limited to approximately 2,000 A/cm^2 for carrier injection into the semiconductor plasma. Also designs requiring excessively high bias voltages to inject or deplete electron and hole concentrations were rejected.

Paper Outline

This work combines two areas of study: semiconductor physics and wave propagation theory. We are using a modeling program that incorporates classical semiconductor theory, Poisson's equation and the continuity equations, to model semiconductor optical waveguiding structures in silicon under various biasing schemes. The paper is separated into four sections:

- Test cases: MOS diode.
- Single injection Transverse MOSFET Analysis.
- Double injection Transverse MOSFET Analysis.
- Double Injection Longitudinal DIFET Analysis.

A more detailed report on this work including PISCES IIB theory of operation, semiconductor theory, optical theory and more test cases can be found in the thesis by the present author (Worcester Polytechnic Institute, 1989).

[13] Private correspondence to Prof. Lionel Friedman, May 11, 1988.

CHAPTER 2

TEST CASE: MOS DIODE

2.1 MOS diode

A MOS diode has a number of advantages as a waveguiding structure. Since the gate is insulated, there is no current through the device and it can be constructed on any type of media with a refractive index less than silicon. However, the waveguiding region will be a very thin area because an insulated gate will not be able to deplete electrons or holes outside of the calculated depletion width. The same restriction also applies when the MOS diode is in accumulation; large concentration changes are possible only in thin (submicron) layers beneath the insulated gate region. Since our work is concerned with single mode optical waveguides, thin waveguides are not a restriction.

Sufficient bias on the gate electrode of the MOS diode will attract and deplete the electrons and holes beneath the insulating layer. If the concentration changes are great enough, then the index of refraction will be altered.

The maximum refractive index change is the peak change in the entire waveguide. It does not necessarily indicate that the device is suitable for optical waveguiding but it is a good indication as to whether large concentrations can be modulated within a device structure. The effective refractive index change, Δn_{eff}, averages the refractive index changes over a particular area of the device using the cosine overlap integral. It is an effective method for comparing two dimensional waveguides.

Based on this simulation, a third order index (10^{-3})change can be expected when a MOS diode is biased to either + or - 25.0 volts from equilibrium.

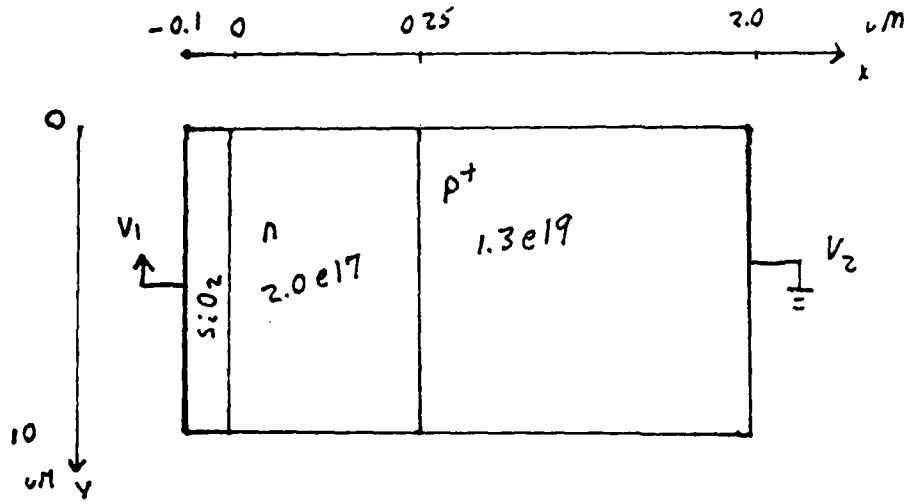
2.1.1 Device Description

The device was simulated using a very simple grid because, if fringe effects are neglected, the device is a one dimensional structure. The solution would not have been more accurate if more points were added.

The entire device is 10 uM wide and 2.1 uM deep. The top 0.1 uM is oxide. Under the oxide layer is a 0.25 uM strip of silicon uniformly doped $2.0x10^{17}$ cm^{-3} p type. The rest of the device structure is highly doped ($1.3x10^{19}$ cm^{-3}) p+ type.

The electrodes are neutral ohmic contacts. Since there is essentially no current through an insulated structure like a MOS diode, the solution was found using a Gummel method. It was solved only for holes since the electrons are in such small quantity.

Figure 1: MOS Diode Diagram



2.1.2 Program Summary

The following PISCES IIB programs were run to characterize the device:

- CPW10 Define device and solve initial solution.
- CPW11 Forward bias gate electrode from 0 to 25.0 volts.
- CPW12 Reverse bias gate electrode from 0 to -25.0 volts.
- CPW13 find voltage, electron and hole concentration when v1= 25.0, v2=0.
- CPW14 find voltage, electron and hole concentration when v1=-25.0, v2=0.
- CPW15 find voltage, electron and hole concentration when v1= 1.0, v2=0.
- CPW16 find voltage, electron and hole concentration when v1= -1.0, v2=0.

See contour plot CPW10_THREED_DOPING for contour plot dimensions.

2.2 MOS diode Theory

Using Sze's Semiconductor Devices as a guide, we can calculate the voltage needed for inversion from the initial doping concentration. [1]

$$\Psi_s(inv) \simeq 2\frac{k_b T}{q}\ln(\frac{N_a}{n_i}) \quad 2.1$$

where

$\Psi_s(inv)$ is the inversion voltage.

$\frac{k_b T}{q}$ is 0.0258 eV at 300 'K.

N_a is the impurity concentration cm^{-3}.

n_i is the intrinsic concentration cm^{-3}.

[1] S M. Sze, Semiconductor Devices Physics and Technology, (Wiely & Sons, New York, 1985), p. 191.

For a doping concentration of $N_a = 2.0x10^{17}$ cm^{-3}, the voltage needed for surface inversion is:

$$\Psi_s(inv) = 2 * 0.0258 * \ln(\frac{2.0x10^{17}}{1.45x10^{10}}) \quad 2.2$$

$$\Psi_s(inv) = 0.8483\ eV \quad 2.3$$

The calculated depletion width is approximately, [2]

$$W_m = \sqrt{\frac{2\epsilon_s\Psi_s(inv)}{qN_a}} \quad 2.4$$

where

W_m is the maximum depletion width.

ϵ_s is the permittivity 11.9 * 8.85x10^{-14} F/cm.

Solving for Wm:

$$W_m = \sqrt{\frac{2 * 11.9 * 8.85x10^{-14} * 0.8483}{1.602x10^{-19} * 2.0x10^{17}}} \quad 2.5$$

$$W_m = 0.0747\mu M. \quad 2.6$$

2.3 Equilibrium

Usually a contour plot of the electron concentration was not made because the electron concentration is so small (except for the forward bias case). Contour plot CPW10_THREED_P is the plot of the hole concentration in the device. The lower doping contours on the left are the oxide layer/p boundary and the higher doping contours are the p/p+ boundary. The gap between the two is the lower doped waveguiding region. CPW10_N and CPW10_P are the three dimensional plots of the electrons and holes while CPW10_V is the plot for the voltage potential at equilibrium. The electron and hole concentrations in cm^{-3} are logarithmic in the z axis and linear in μM along x and y directions. From the hole plot, CPW10_P, the region along the y axis is the 0.1 uM oxide layer and the "stair" is the lower doped 0.25 uM p region. Note on the electron plot, CPW10_N that there is a slightly higher concentration of electrons under oxide layer. This is the result of the built-in voltage.

2.4 Positive Biasing of Gate Electrode

CPW15_V is a three dimensional plot of the voltage potential when the gate is biased to one volt. A small applied voltage depletes holes without attracting a large concentration of electrons.

The contour plot of the holes, CPW15_THREED_P, shows that the hole concentration has changed from $1.0x10^{17}$ to $1.0x10^{16}$ under the oxide layer. However, this is only a very thin strip (less than 0.025 uM). The change in hole concentration induces a maximum refractive index change of 2.48e-4 at 1.3 uM and 3.51e-4 at 1.55 uM when the gate is biased to +1.0 volts.

[2] Ibid.,.

A positive forward bias of 25.0 volts (CPW13_V) attracts a large concentration of electrons (CPW13_THREED_N) under the oxide layer. Compared with equilibrium (CPW10_THREED_P), the hole concentration is shifted away from the oxide so that a low density of holes (CPW13_THREED_P) is left.

At 25.0 volts, the increased electron concentration causes a maximum index change of 2.186e-3 at 1.55 uM (CPW13_INDEX_155) and 1.545e-3 at 1.3 uM (CPW13_INDEX_13). However, the index change created from the accumulation of electrons is somewhat negated by the depletion of holes. Graph CPW13_INDEX_13_5 is a one dimensional slice of the index change vs distance into the MOS diode. The cpw13_13 line shows the region under the oxide layer with the extra electrons but it also highlights a secondary effect; the accumulation of holes away from the oxide region. This would cause a slight increase in the refractive index and results in even better light confinement.

2.5 Reverse Biasing of the Gate Electrode

The three dimensional plot CPW16_V is the plot of the voltage potential with the gate negatively biased to -1.0 volt. Since a negative voltage repels electrons, the electron concentration is negligible. Comparing the two hole contour plots (CPW10_THREED_P at equilibrium and CPW16_THREED_P at -1.0), there is only a slight difference between them. Examining the data files, the hole concentration changes from $1.4397x10^{17}$ cm^{-3} to $2.8119x10^{17}$ cm^{-3} under the oxide layer.

The index change is slightly larger for accumulation than depletion. The contour plots CPW16_INDEX_13 and CPW16_INDEX_155 are contour plots of the refractive index change when the gate bias is changed from 0 to -1.0 volts and the maximum index change at 1.55 uM is -4.36e-4 and -3.08e-4 at 1.3 uM.

Reverse biasing the gate to -25.0 volts (CPW14_v) causes a large increase in the hole concentration under the oxide layer. Unlike the depletion case, the electron concentration remains minimal.

A contour plot of the holes, CPW14_THREED_P, shows an increase to $1.0x10^{18}$ cm^{-3} of holes under the oxide layer. The maximum predicted index change at 1.3 uM is -4.688e-3 and -6.641e-3 at 1.55 uM.

Graph CPW15_INDEX_13_5 is a summary of the change in refractive index vs distance in the MOS diode due to attraction or depletion. Note that attraction is a larger effect and that the only index change is directly under the oxide layer. Combining the two effects, attraction and depletion, two index change contour plots were created (CPW17_INDEX_13 and CPW17_INDEX_155) with a gate voltage changed from -1.0 volt to +1.0 volt; the idea being that the index change from the depletion of holes and attraction from the holes would add together. However, this is not the case because the index change based on the change in hole concentration is not a linear effect but depends on a factor of 0.8. There is an increase but not as large as hoped.

2.6 Results and Discussion

Below is a summary of maximum predicted refractive index change due to change in gate voltage.

Table 1: Summary of Peak index change vs change in gate voltage for MOS Diode

title	Initial gate	Final gate	$\Delta n(1.3)$	$\Delta n(1.55)$
CPW13	0.0	25.0	-1.545e-3	-2.186e-3
CPW14	0.0	-25.0	-4.688e-3	-6.641e-3
CPW15	0.0	1.0	0.248e-3	0.351e-3
CPW16	0.0	-1.0	-0.308e-3	-0.436e-3
CPW17	-1.0	1.0	0.484e-3	0.686e-3

Based on the predicted electron and hole changes from the PISCES IIB simulations, one can expect third order index changes from a MOS diode when the bias is changed from 0 to +/- 25.0 volts. Both a positive or negative 25.0 gate voltage induce a negative index change. A pure depletion effect, where the electrons are depleted beneath the gate without a significant accumulation of holes, was not pursued because the largest index possible in depletion is from $2.0x10^{17}$ cm^{-3} holes; still a -4 order result.

From the PISCES IIB analysis of the MOS diode, one can draw the following conclusions:

- Large index changes are possible when the gate voltage is above the inversion voltage.
- Accumulation (negative gate voltage) seems to be the larger effect because the electron and hole concentrations changes do not cancel each other out.
- A slightly larger index change from depletion would result if the MOS structure was n type instead of p type. Then negative biasing would attract holes and deplete electrons. Since holes cause a larger index change, the predicted index change would be larger.
- The push-pull effect, changing from a positive to a negative voltage, seems to improve the index change. However, if the index change is based on the change in the hole concentration then it is not a linear change.
- The index change region is very small, less than 0.1 uM. Increasing the p region doping would decrease the depletion width and make that region even smaller.
- Decreasing the doping would mean that it would take an even higher voltage to get a hole concentration of $1.0x10^{18}$ under the oxide layer.
- The MOS diode has the advantage of using an applied electric field to induce electron and hole changes so that current density and power dissipation are not limiting consideration.

CHAPTER 3

SINGLE AND DOUBLE INJECTION TRANSVERSE SILICON MOSFET WAVEGUIDE STRUCTURES

3.1 Introduction

Based on the papers presented on waveguide modulators at the recent SPIE conference in Boston, [1] a silicon MOSFET device is a novel approach for phase modulating light signals. Richard Soref's designs for MOSFET optical structures were used as a guide to create single and double injection MOSFET structures. [2] A single injection structure injects only holes or electrons into the waveguiding region while a double injection device injects both holes and electrons. Above the middle of the device is an oxide rib which forms an insulated gate contact. A voltage potential on the insulated gate modulates I_{ds}, the drain to source current, and the charge distribution. The result is a phase modulator in the xy plane such that light propagating along the z axis (into the page) will be phase modulated by changes induced in the refractive index of the waveguiding region.

Unlike the longitudinal DIFET ‡ structures described later in this report, current density is a limiting factor because the drain and source contacts are separated only by the waveguide width. The distance between contacts is an important design consideration since, for a given biasing condition, the current density is inversely proportional to the separation between the drain and source contacts. The current flow from the drain-source is perpendicular to the direction of light propagation. Since both the gate and substrate contacts are insulated, there are no electrons or holes swept out by either contact. An analysis of a MOSFET also differs from a DIFET because short channel effects must be taken into consideration.

Using the PISCES IIB semiconductor simulation program, the electrical behavior of the MOSFET can be analyzed. The electron and hole concentrations are then used to calculate the predicted optical properties of the device.

3.1.1 Device Description

A number of different MOSFET geometries were simulated. The width of the simulated device (6 or 14 uM) depends on whether a 2 uM or a 10 uM insulated gate was simulated. The MOSFETs are 4.1 uM thick and the top 0.1 uM above the waveguide region is the oxide layer between the waveguiding region and the gate electrode. The 0.1 uM above the drain and source regions is an insulating material, such as air, with an index of refraction of 1.0. The gate contact is centered on the device and the drain and source regions are 0.5 uM wide and 0.5 uM deep, 1.0 uM from the gate. The channel region is 2.0 uM deep along the whole

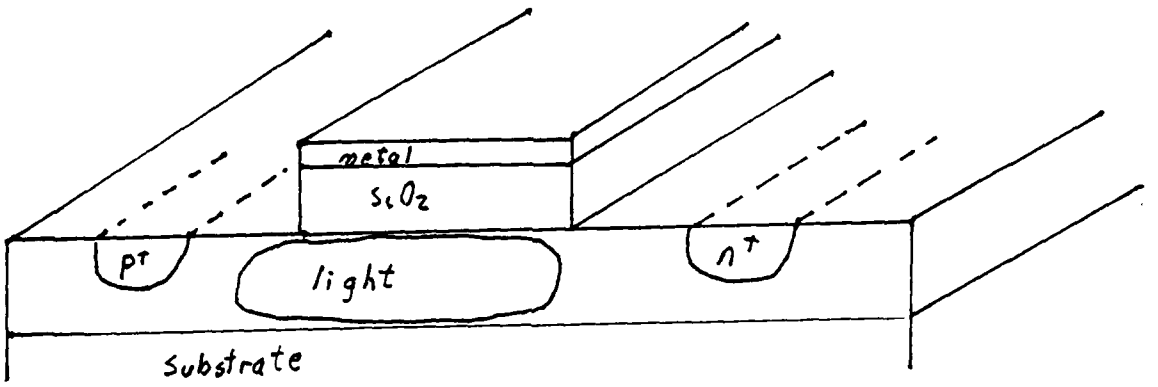
[1] SPIE Conference on Fiber Opitcs, Optoelectronics and Laser Applications, vol. 985 to 999, Boston, MA, September 6-10, 1988.

[2] Richard Soref, "Notes on Silicon Triodes", private memo, January 19, 1988.

‡ Double Injection Field Effect Transistor

device and the entire structure is on top of a 2.0 uM oxide layer. The waveguiding region is bounded top and bottom by the insulated gate and oxide substrate. The two oxide layers will insure that the lightwave is strongly confined and symmetric in the y direction. Along the x direction, the waveguide will be bounded on either side by the heavily doped drain and source regions. the dopant regions should decrease the refractive index by ~0.001; sufficient for light confinement.

Figure 2: MOSFET waveguide structure



The drain is doped $1.3x10^{19}$ cm^{-3} p-type and the source is doped $1.3x10^{19}$ cm^{-3} either n+ or p+ depending on whether single or double injection was used. The waveguide doping concentration ranged from $1.0x10^{14}$ to $5.0x10^{17}$ cm^{-3} depending on the program. The electrodes are neutral ohmic contacts except for the gate which is aluminum. Simulations were run without recombination effects (ideal) and with Schockley-Reed-Hall (SRH) recombination. In this way, the effects of recombination can be fully analyzed.

Below is a summary of the MOSFETs simulated.

Table 2: Summary of MOSFET device Structures

title	source (cm^{-3})	channel (cm^{-3})	gate width (μM)	drain (cm^{-3})	recombination
MDF50	$1.3x10^{19}$ p+	$1.0x10^{14}$ n	2.0	$1.3x10^{19}$ p+	
MDF70	$1.3x10^{19}$ p+	$1.0x10^{14}$ n	10.0	$1.3x10^{19}$ p+	
MDF90	$1.3x10^{19}$ p+	$5.0x10^{17}$ n	2.0	$1.3x10^{19}$ p+	
MDF100	$1.3x10^{19}$ p+	$1.0x10^{14}$ n	2.0	$1.3x10^{19}$ p+	
MDF120	$1.3x10^{19}$ n+	$1.0x10^{14}$ n	10.0	$1.3x10^{19}$ p+	
MDF140	$1.3x10^{19}$ n+	$1.0x10^{14}$ n	10.0	$1.3x10^{19}$ p+	SRH
MDF200	$1.3x10^{19}$ p+	$1.0x10^{15}$ n	10.0	$1.3x10^{19}$ p+	
MDF220	$1.3x10^{19}$ p+	$1.0x10^{15}$ n	10.0	$1.3x10^{19}$ p+	SRH

3.1.2 Early Test Cases

MOSFETS MDF50, MDF70, MDF90 and MDF100 are early test cases that will not be discussed in great detail. Although, these simulations did not include recombination effects, they were used to draw initial impressions about the waveguiding possibilities of a MOSFET device. In particular:

- There is a trade-off between the concentration of holes attracted to the gate region and current density of the device. In order to create a large index change, concentration changes should be on the order of $1.0x10^{18}$ cm^{-3}. However, the current density exceeds 2000 A/cm^2 before that injection level.
- Comparing the various biasing schemes, one would expect that the larger voltage differences would create the larger index changes. However, this is not always true because a positive bias on the gate electrode, will not only deplete the waveguide of holes, but also attract a high concentration of electrons. The two effects, depletion of holes and accumulation of electrons, cancel out the expected double injection change in the refractive index.
- For a single injection MOSFET, a wider gate region (10 uM vs 2 uM) produces the same refractive index change at a lower current density. Therefore, higher refractive index changes are possible if the device has a high gate width to waveguide depth ratio.
- High initial waveguide doping concentrations will not be suitable for electron and hole injection.
- For an insulated gate waveguide structure, the effective index change is not significantly altered when the waveguide is 2.0 uM thick. However, if we only consider the 0.1 uM under the gate region then the effective index changes vs gate voltage will be significant.

3.2 Index Change Vs. Electron and Hole Distribution Changes

The index of refraction has both real and imaginary components: [3] .

$$N = n + i\kappa \quad 3.1$$

The extinction coefficient, κ, is defined: [4]

$$\kappa = \frac{\alpha\lambda}{4\pi} \quad 3.2$$

where

α is the absorption coefficient.

λ is the light wavelength.

[3] T. S. Moss, Optical Properties of Semiconductors, (London, Butterworths Scientific Publications, 1959), p. 2

[4] Solomon Musikant, Optical Properties, (Marcel Dekker, New York, 1985), p. 7

In silicon, changes in the real component of the refractive index, Δn, can be calculated using data from Richard Soref. [5] . For analysis, the electron and hole concentrations at each xy data point in the PISCES IIB mesh are compared with concentrations under a different biasing condition. The electron and hole difference at each xy point is put into an index change data table and multiplied by the index change constant. Two refractive index change plots are necessary because silicon has different optical constants at $\lambda = 1.3$ uM and $\lambda = 1.55$ uM. [6] At 1.3 uM wavelength, the predicted refraction effect is

$$\Delta n_e(1.3) = -6.2x10^{-22}\Delta N_e \quad 3.3$$

$$\Delta n_h(1.3) = -6.0x10^{-18}(\Delta N_h)^{0.8} \quad 3.4$$

while at 1.55 uM wavelength,

$$\Delta n_e(1.55) = -8.8x10^{-22}\Delta N_e \quad 3.5$$

$$\Delta n_h(1.55) = -8.5x10^{-18}(\Delta N_h)^{0.8} \quad 3.6$$

There is a linear relationship between the index of refraction and the electron concentration but a nonlinear relationship between Δn and the hole concentration. Since the index change constants are negative, electron or hole *injection* will *decrease* the refractive index while *depletion* will *increase* the refractive index.

Finally, the refractive index changes due to the change in electron concentration and change in hole concentration at each xy mesh point are added together to form the total refractive index change profile.

$$\Delta n = \Delta n_e + \Delta n_h \quad 3.7$$

There are two methods to compare the refractive index information. One is the maximum index change, Δn_{max}, vs. gate voltage at 1.3 uM and 1.55 uM wavelengths. This is the peak refractive index change calculated within the device due to a change in the applied gate voltage. The maximum index change is a good measure of how well charge can be manipulated in a device structure but is not necessarily an indication how well a device will act as a phase modulator.

The second method is to calculate the effective index change, Δn_{eff} which averages the index changes over the waveguiding area. The effective refractive index change is defined: [7] .

$$\Delta n_{eff} = \frac{\int_{x_l}^{x_r}\int_{y_t}^{y_b}|\Psi(x,y)|^2\partial n(x,y)dxdy}{\int_{x_l}^{x_r}\int_{y_t}^{y_b}|\Psi(x,y)|^2dxdy} \quad 3.8$$

where

$\Psi(x,y)$ is the electric field distribution.

$\partial n(x,y)$ is the change in the refractive index.

[5] Richard Soref and Brian Bennett, "Electrooptical Effects in Silicon", IEEE Journal of Quantum Electronics, Vol QE-23, January, 1987, p. 127

[6] Private correspondence to Prof. Lionel Friedman, May 11, 1988.

[7] M. J. Adams, S. Ritchie and J. J. Robertson, "Optimum Overlap of Electric and Optical fields in Semiconductor Waveguide Devices", Applied Physics Letters, March 31, 1986, p. 820

The electric field distribution inside the waveguide is a product of cosines. If only single mode waveguides are considered then equation electric field can be expressed as

$$\Psi(x,y) = C_l \cos(\frac{|x_0 - x|}{d_x} U_x) \cos(\frac{|y_0 - y|}{d_y} U_y) \quad 3.9$$

where

C_l is the normalization constant,

x_0,y_0 are the center of the waveguide,

d_x,d_y are the length/2,width/2 of the waveguide,

U_x is the boundary value calculated from the refractive index at the x boundaries for a given wavelength,

and U_y is the boundary value calculated from the refractive index at the y boundaries for a given wavelength.

The integral limits are the defined boundaries of the waveguide in the x and y direction. For a MOSFET, the limits in the x direction are the edges of the source and drain regions at 1.0 uM and 13.0 uM and the y axis boundaries are the oxide layers.

3.3 Optical Waveguide Dimensions

The numerical aperture, NA, is a measure of the light gathering power of a waveguide, defined by [9]

$$NA = \sqrt{n_1^2 - n_2^2} \quad 3.12$$

where

n_1 is the refractive index of the waveguide region,

and n_2 is the refractive index of the outer boundary region.

A waveguide can support a number of modes depending on the wavelength of the light, numerical aperture of the waveguide defined: [10]

$$V = \frac{2\pi a}{\lambda} NA \quad 3.13$$

where

V is the normalized waveguide width or normalized frequency,

λ is the wavelength,

and $2a$ is the guide width.

[9] Y. Suematsu and Ken-Ichi Iga, Introduction to Optical Fiber Communication, (Wiley & Sons, New York, 1982), p. 18

[10] Ibid., p. 23

In a two dimensional waveguide, the maximum number of supported modes, N_m is defined: [11]

$$N_m \simeq \frac{1}{2}V^2 \quad 3.14$$

The cut-off wavelength is defined as the wavelength, λ_0, at which any shorter wavelength results in a multimode waveguide. [12] The cut-off condition is the point when the waveguiding region and cladding boundary no longer maintains total internal reflection. At 1.3 uM, the maximum waveguide diameter for single mode operation can be calculated.

Using the high frequency dielectric constant for silicon, 11.8, and taking the square root, a silicon waveguide would have a refractive index of n1 = 3.4351. [13] Assuming an index change, Δn, of 0.001 and using equations 3.12, 3.13, and 3.14, then the maximum waveguide size can be calculated.

The numerical aperture is found to be:

$$NA = \sqrt{3.43610^2 - 3.4351^2} \quad 3.15$$

$$NA = 0.0824 \quad 3.16$$

At λ = 1.3 uM,

$$2.405 > V = \frac{2\pi a}{\lambda} NA \quad 3.17$$

$$2a < \frac{2.405 * 1.3\ uM}{\pi * 0.0824} \quad 3.18$$

$$2a < 12.078\ uM \quad 3.19$$

Anything greater than 12.00 uM will result in a multimode waveguide at the 1.3 uM wavelength.

3.4 Analysis Method

1. Design MOSFET structure using PISCES IIB semiconductor simulation programs.
2. Analyze the electrical characteristics of the semiconductor.
3. Gather data on electron, hole and current density distribution for different biasing conditions.
4. Calculate the change in the refractive index at each mesh point.
5. Use the overlap integral to calculate the effective index change in the waveguide region of the device.

[11] Dietrich Marcuse, Theory Of Dielectric Optical Waveguides (Academic Press, New York, 1974) p. 76

[12] Lynn Hutcheson, Integrated Optical Circuits and Components, (Marcel Dekker, New York, 1987), p. 19

[13] Charles Kittel, Introduction to Solid State Physics (Wiley & Sons, New York, 1986), p. 207

CHAPTER 4

SINGLE INJECTION TRANSVERSE MOSFET OPTICAL WAVEGUIDE

4.1 Introduction

A single injection MOSFET is a simpler design than a double injection MOSFET because there is only one injected carrier, either electrons or holes, that can alter the refractive index of the waveguide. In a transverse design, the light ray travels in the z direction into the page. The waveguide is in the xy plane perpendicular to the incoming light ray.

Model MDF200 is evaluated without including any recombination effects while model MDF220 includes Schockley-Reed-Hall recombination. The same set of programs were run to characterize the two models so only the differences between MDF200 and MDF220 will be mentioned.

This model was designed with the following assumptions:

- The heaviiy doped regions are uniform abrupt junctions that are 0.5 uM deep.
- Both the drain and source regions are doped $1.3x10^{19}$ cm^{-3} p+ for a single injection device.
- The gate and substrate are insulated for stronger light confinement.
- The maximum waveguide width is 12.0 uM wide to insure that only a single mode will propagate.

4.2 MOSFET Semiconductor Design

The single injection MOSFET is similar to a p-channel MOSFET; a sufficient negative gate voltage will create an inversion layer between the p+ drain and source regions. The substrate is an oxide layer with a neutral ohmic contact. This contact would not be necessary for a device with an insulated substrate and is not used to bias the MOSFET.

The voltage at the neutral contacts along the p+ drain and source is calculated from the doping:

$$V_{contact} = \phi \quad \begin{matrix} + \\ - \end{matrix} \frac{k_b T}{q} \ln\left(\frac{N_a}{n_i}\right) \qquad 4.1$$

where

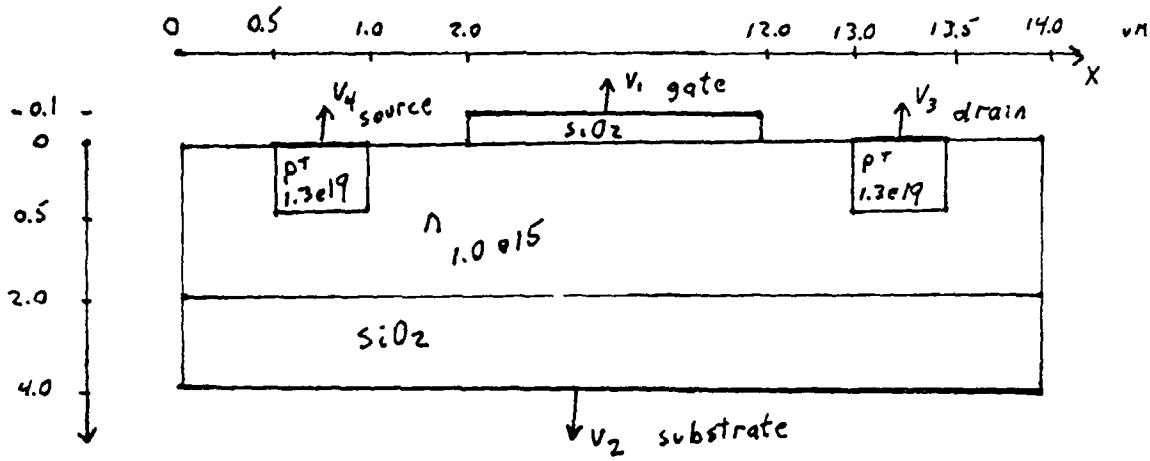
ϕ is the applied contact voltage,

$\frac{k_b T}{q}$ equals 0.0258 eV at 300 K,

N_a is the doping concentration,

n_i is the intrinsic concentration,

Figure 3: Single Injection MOSFET for MDF200 and MDF220



and +/− depends on the doping; + for n type.

$$V_{contact} = -0.0258\ eV \ln\left(\frac{1.3x10^{19}}{1.45x10^{10}}\right) \quad 4.2$$

$$V_{contact} = -0.59271 \quad 4.3$$

The gate voltage is calculated using the work function for aluminum. From equation 2.42 of the PISCES IIA manual, the equation for a Schottky contact on silicon is:

$$V_{contact} = \chi + \frac{E_g}{2} + \frac{k_b T}{2q} \ln\left(\frac{N_c}{N_v}\right) - \phi_m - V_{applied} \quad 4.4$$

where

χ is the electron affinity, 4.17 eV,

$\frac{E_g}{q}$ is the energy gap, 1.08 eV,

N_c, N_v are the electron and hole concentrations,

ϕ_m is the work function of the metal, 4.17 eV,

and $V_{applied}$ is the applied voltage.

The values for silicon are from the PISCES IIB MATERIALS simulation card. For an insulated contact, such as aluminum on an oxide layer, equation 1.4 reduces to

$$V_{contact} = \chi + \frac{E_g}{2} - \phi_m - V_{applied} \quad 4.5$$

$$V_{contact} = 0.54\ eV \quad 4.6$$

If the junction depth is limited to 0.5 uM, then the maximum waveguide doping can be calculated. [1]

$$W_m = \sqrt{\frac{2\epsilon_s(2\frac{k_bT}{q})\ln(\frac{N_a}{n_i})}{qN_a}} \quad 4.7$$

where

W_m is the maximum depletion width,

and ϵ_s is the permittivity.

Assuming a depletion width of 0.5 uM and solving equation 1.7,

$$0.5\ uM = \sqrt{\frac{4 * 11.9 * 8.85x10^{-14} * 0.0258\ln(\frac{N_a}{1.45x10^{10}})}{1.602x10^{-19}N_a}} \quad 4.8$$

Simplifying

$$e^{3.685x10^{-15}N_a} = \frac{N_a}{1.45x10^{10}} \quad 4.9$$

Taking the derivative and solving, the maximum waveguide doping is $2.67x10^{15}$ cm^{-3}. However, when the V_{ds} is applied, the electron concentration will increase and the maximum depletion width will decrease. But a waveguide doping concentration of $1x10^{15}$ cm^{-3} is a good approximation.

The built-in voltage is defined as [2]

$$V_{bi} = \frac{k_bT}{q}\ln(\frac{N_aN_d}{n_i^2}) \quad 4.10$$

If the electron concentration in the waveguide is $1x10^{15}$ and the doping regions are $1.3x10^{19}$ p+ then the built-in voltage is

$$V_{bi} = 0.0258 * \ln(\frac{1x10^{15}1.3x10^{19}}{2.08x10^{20}}) \quad 4.11$$

$$V_{bi} = 0.819\ eV \quad 4.12$$

The surface potential at the semiconductor-oxide interface is [3]

$$\Psi_s = \frac{qN_aW^2}{2\epsilon_s} \quad 4.13$$

[1] S. M. Sze, Semiconductor Devices Physics and Technology, (Wiley & Sons, New York, 1985), p. 191

[2] Ben Streetman, Solid State Electronic Devices, (Prentice-Hall, Englewood Cliffs, NJ, 1980), p. 140

[3] Sze, p. 191

If the effective oxide charge is not taken into consideration then the voltage needed for surface inversion is twice the midgap potential or:

$$\Psi_s(inv) = 2\frac{k_b T}{q}\ln(\frac{N_a}{n_i}) \quad 4.14$$

$$\Psi_s(inv) = 2 * (0.0258)\ln(\frac{1x10^{15}}{1.45x10^{10}}) \quad 4.15$$

$$\Psi_s(inv) = 0.626\ eV \quad 4.16$$

Since this a transverse waveguide limited to single mode operation, there is a limit on the gate width. Therefore, we need to take short channel effects into consideration. The maximum channel length is the depletion width of the source region plus the depletion width of the drain region. [4]

$$L = W_s + W_d \quad 4.17$$

If the source contact is at ground potential then the maximum V_{ds} can be calculated. Using equations 1.1 and 1.9

$$L = \sqrt{\frac{2\epsilon_s}{qN_a}V_{bi}} + \sqrt{\frac{2\epsilon_s}{qN_a}(V_{bi} + V_d)} \quad 4.18$$

If the channel length is 12.0 uM and the built-in voltage is 0.819 eV then the maximum V_{ds} is

$$12.0\ uM = 3.16\ uM + \sqrt{1.31(0.819 + V_{ds})} \quad 4.19$$

Solving for V_{ds} the maximum voltage is over 58.0 volts if the channel length is 12.0 uM. In this instance, short channel effects are not significant. However, if the gate width is narrowed, short channel effects may become more important. With this model, the minimum channel length can be calculated. If V_{ds} is zero, then the minimum L is 2.18 uM.

4.2.1 Program Summary

The following PISCES IIB programs were run to characterize the device:

- MDF220 Define device and solve initial solution without recombination.
- MDF221 Increase V_{ds} to 3.0 volts and create I-V curves for v1 = 0, v2 = 0, v3 from 0 to 3.0 volts and v4 = 0.
- MDF222 find electron, hole, voltage and current density profiles for v1 = 0, v2 = 0, v3 = 3.0 and v4 = 0.
- MDF223 Increase V_{ds} to 5.0 volts: v1 = 0, v2 = 0, v3 = 5.0 volts and v4 = 0
- MDF224 find electron, hole, voltage and current density profiles for v1 = 0, v2 = 0, v3 = 5.0 volts and v4 = 0.
- MDF225 Decrease gate voltage (v1) from 0 to -5.0 volts with V_{ds} = 3.0 volts.
- MDF226 Increase gate voltage (v1) from 0 to 5.0 volts with V_{ds} = 3.0 volts.
- MDF227 find electron, hole, voltage and current density profiles for v1 = -5.0, v2 = 0, v3 = 3.0 and v4 = 0.

[4] Ibid., p 213.

- MDF228 find electron, hole, voltage and current density profiles for v1=5.0, v2=0, v3=3.0 and v4=0.
- MDF229 find electron, hole, voltage and current density profiles for v1=-0.5, v2=0, v3=3.0 and v4=0.
- MDF231 find electron, hole, voltage and current density profiles for v1=0.0, v2=0, v3=3.6 and v4=0.
- MDF232 increase gate voltage (v1) from 0 to 2.0 volts with V_{ds} = 3.6 volts.
- MDF233 find electron, hole, voltage and current density profiles for v1=2.0, v2=0, v3=3.6 and v4=0.
- MDF234 find electron, hole, voltage and current density profiles for v1=2.8, v2=0, v3=3.6 and v4=0.

The program name is used as the prefix to identify the three dimensional plots, contour plots and I-V curves. Note that the contour plots are not plotted to scale because the contour algorithm skews the dimensions.

4.3 Single Injection MOSFET: MDF220

Contour plots MDF220_THREED_P and MDF220_THREED_N are the equilibrium hole and electron distributions. The plots are symmetrical with the highly doped p+ regions on either side of the waveguide.

Below is a summary of the different gate bias voltages used to analyze the MDF220 model.

Table 3: Summary of Bias Voltages for MDF220 Single Injection MOSFET

title	gate (v1)	substrate (v2)	drain (v3)	source (v4)
MDF220	0.0	0.0	0.0	0.0
MDF222	0.0	0.0	3.0	0.0
MDF227	2.0	0.0	3.0	0.0
MDF228	-1.0	0.0	3.0	0.0
MDF231	0.0	0.0	3.6	0.0
MDF233	2.0	0.0	3.6	0.0
MDF234	2.8	0.0	3.6	0.0
MDF235	0.0	0.0	1.0	0.0
MDF238	-4.4	0.0	1.0	0.0

4.4 MDF220 Drain-source Biasing

Three different biasing schemes were simulated to find the greatest effective index change in the waveguiding region. First, a V_{ds} voltage was applied and the gate voltage was used to modulate the hole concentration; negative gate voltage increases the hole concentration and positive voltage decreases the hole concentration. In this case, most of the holes are injected from the drain and source contacts. Next the maximum V_{ds} voltage was applied to inject holes and a positive gate voltage was used only to deplete the holes. Finally a small V_{ds} voltage, which injects very few holes when the gate is zero, was applied and a negative gate voltage was used to attract holes.

Graph MDF221_IV is a plot the drain-source current vs V_{ds} applied voltage. Curve MDF201 is the ideal I_{ds} current without recombination and the MDF221 curve includes recombination. From this graph, including recombination effects seems to make the most difference at lower currents for single carriers simulations. As the V_{ds} increases there is less difference between the two currents.

4.5 MDF220 Analysis with V_{ds} = 3.0 volts

Unlike the double injection device, a larger V_{ds} voltage is needed to inject holes into the waveguiding region. Almost all of the holes are injected directly beneath the gate while the electron concentration throughout the device is not enough to influence the index of refraction. Graph MDF226_IV is a plot of the drain-source current vs applied gate voltage.

Unlike the double injection device, an applied gate voltage will modulate the drain-source current. From graph MDF226_IV, a negative gate voltage will increase the drain-source current while a positive voltage will pinch off the current. Curve MDF200 on graph MDF226_IV is the idealized plot without including recombination effects. When the gate voltage is positive, at lower currents, the idealized current is lower than the MDF220 model. However, at higher currents the two curves are similar.

4.5.1 Positive Biasing of MDF220 Gate with V_{ds} = 3.0 volts

If the gate is biased to +2.0 volts, the I_{ds} current is pinched off. The holes, injected from the drain-source contacts, are depleted (MDF227_THREED_N). At +2.0 volts, the electrons are starting to collect under the gate but not in sufficient numbers to affect the index of refraction. If the gate voltage is increased enough, then the depletion of holes would be offset by the accumulation of electrons.

The current density is negligible with this biasing scheme.

4.5.2 Negative Biasing of MDF220 Gate with V_{ds} = 3.0 volts

A negative gate voltage will attract holes underneath the gate region to further increase the hole concentration. If the negative gate voltage is increased, it will create an inversion layer, a p-channel between the p+ drain and source, under the gate oxide layer. If the gate voltage is -1.0 volts, then the peak current density is over 2,000 A/cm^2, (MDF228_|JTOTAL|). The peak hole concentration (MDF228_THREED_P), while increased from MDF222 (V_g=0), is still not high enough to induce Δn=0.001 index changes in the waveguiding region.

Contour plot MDF228_INDEX_155 is the predicted index changes at λ = 1.55 uM wavelength when the gate voltage is changed from 0 to -1.0 volts. The maximum index changes are -0.96e-4 at λ = 1.3 uM and -1.32e-4 at λ = 1.55 uM.

4.6 MDF220 Analysis with V_{ds} = 3.6 volts

If the V_{ds} voltage is increased to 3.6 volts, then the peak current density is almost 2,000 A/cm^2, (MDF231_|Jtotal|). Any applied negative bias on the gate electrode would increase the current density above the limit for stable operation. The hole concentration, MDF231_THREED_P, under the gate is increased to $1x10^{16}$ cm^{-3} but the electrons under the gate are depleted (MDF231_THREED_N).

4.6.1 Positive Biasing of MDF220 Gate with V_{ds} = 3.6 volts

A positive gate voltage will pinch off the I_{ds} current as graph MDF232_IV demonstrates. When the V_{ds} is increased from 3.0 and 3.6 volts, it requires a slightly larger positive gate voltage to deplete the waveguiding region of holes.

Contour plots MDF233_THREED_N and MDF233_THREED_P are plots of the electron and hole concentrations when the gate is biased to +2.0 volts. Note that the hole concentration has only decreased to $1x10^{15}$ cm^{-3} under the gate. However, If the gate voltage is increased to 2.8 volts, then the holes, MDF234_THREED_P, are depleted and the I_{ds} current is pinched off.

4.7 MDF220 Analysis with V_{ds} = 1.0 volts

If the V_{ds} voltage is 1.0 volts, then there is very little change in the electron and hole concentrations from equilibrium as contour plots MDF235_THREED_N and MDF235_THREED_P indicate. From graph MDF237_IV, a small gate voltage is needed to completely pinch off the I_{ds} current. However, the hole concentration is insignificant under the gate so that a positive gate voltage will not be needed to deplete the holes. The largest increase in I_{ds} is from 0 to -0.6 volts and a larger negative gate voltage does not add significantly to the I_{ds} current.

4.7.1 Negative Biasing of MDF220 Gate with V_{ds} = 1.0 volts

A negative gate voltage of -4.4 volts will attract an increased hole concentration of $1x10^{17}$ cm^{-3} under the insulated gate (MDF239_THREED_P). The electron concentration is still very low throughout the device (MDF239_THREED_N). A small V_{ds} voltage and a large negative gate voltage to attract holes is a superior biasing method to a large V_{ds} voltage. This biasing method has a significant advantage because the current density is much lower. The predicted peak current density from the three dimensional profile MDF239_| Jtotal | is 233 A/cm^2.

The maximum refractive index changes are -3.14e-4 at 1.3 uM and -4.45e-4 at 1.55 uM from contour plots MDF239_INDEX_13 and MDF239_INDEX_155.

4.8 Results and Discussions

The maximum refractive index change is the peak index change calculated from the change in electron and hole concentrations under different biasing conditions. For an insulated structure like a MOSFET, this is typically directly beneath the gate region. It must be kept in mind, however, that just because a model has a high maximum index change, it does not necessarily indicate that it will be a design worthy of future consideration. A better method of comparison is the effective index change, Δn_{eff} where the index changes are averaged over the entire waveguiding region.

In this instance the effective index changes are very small because an applied gate only modulates the 0.1 uM region underneath the gate. If the entire 24 uM^2 of the waveguide is averaged, the effective index changes, shown in the table 'Effective Refractive Changes vs. Gate Voltage for a 1.2 uM^2 Guide', are even smaller.

Table 4: Effective Refractive Index changes vs Gate Voltage for a 1.2 uM² Guide

title	V_{ds}	Initial gate	Final gate	$\Delta N_{eff}(1.3)$	$\Delta N_{eff}(1.55)$
MDF227	3.0	2.0	-1.0	-4.30e-5	-6.10e-5
MDF228	3.0	0.0	-1.0	-4.30e-5	-6.10e-5
MDF233	3.6	0.0	2.0	4.40e-5	6.20e-5
MDF234	3.6	0.0	2.8	4.60e-5	6.50e-5
MDF239	1.0	0.0	-4.4	-1.68e-4	-2.37e-4

Based on the single injection MOSFET simulations, one can draw the following conclusions:

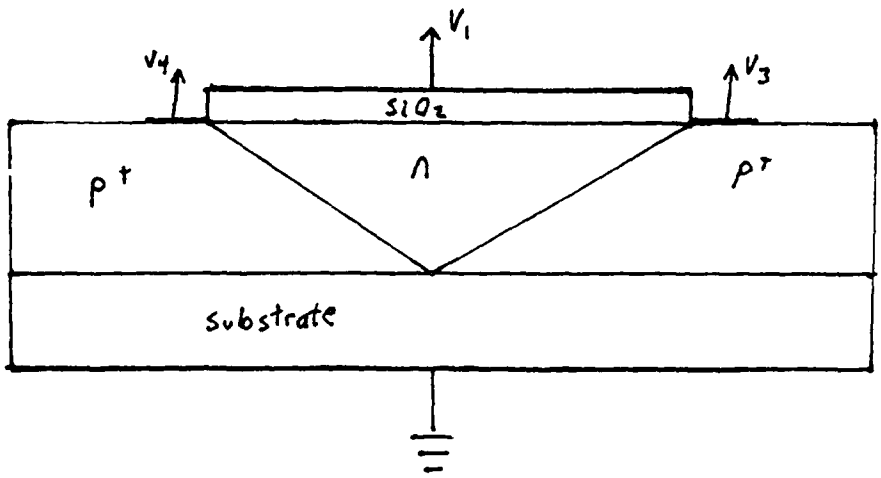
- An applied gate voltage will be able to modulate the drain-source current of a single injection MOSFET.
- Transverse waveguides will be limited by the peak current densities inherent in the design because the drain and source are close together.
- Insulated gate structures, designed to modulate the electron and hole concentrations, will have to be very thin waveguides since the maximum depletion width decreases with increased concentrations.
- The applied gate voltage does not have to completely deplete the waveguide of carriers to induce refractive index changes.
- For single injection PNP structures with uniform abrupt junctions, most of the current will be beneath the gate oxide layer, not distributed throughout the device.
- For single injection MOSFET designs, attracting holes using an applied gate voltage, induces a greater refractive index change than hole injection from the drain and source contacts.

Although the results for the single injection MOSFET were not as great as hoped, there are still possibilities for this geometry. If the drain and source regions were tapered so that the bottom of the doping regions were close together while beneath the gate they were further apart, possibly a larger waveguiding area within the device could be utilized.

A V-grove design would allow a higher applied gate voltage for accumulation of electrons or holes in a larger area. For MOSFET devices with abrupt junctions, the holes accumulate in only a small area beneath the gate.

This geometry would be more difficult to evaluate optically because a simple two dimensional cosine electric field distibution could not be used.

Figure 4: MOSFET with a V groove waveguiding region



CHAPTER 5

DOUBLE INJECTION TRANSVERSE MOSFET OPTICAL WAVEGUIDE

5.1 Introduction

> A two-carrier current will be larger than either one-carrier current in the same crystal. (However), a new limitation on current flow makes its appearance -loss of current carriers through recombination. The injected electrons and holes can mutually recombine before they complete their respective transits between cathode and anode. Normally this recombination is a two-step process which takes place through localized recombination centers ... In steady state, the net rates of electron capture and hole capture by each set of recombination centers must be equal. [1]

Double injection in the silicon MOSFET semiconductor is achieved by changing the source contact of the p+ single injection MOSFET model to n+ type. Compared to the single injection simulation, MDF220, the device dimensions and contacts are the same but the waveguide doping is only $1.0x10^{14}$ cm^{-3} n-type.

A double injection MOSFET will have a greater current density than a single injection MOSFET. Another way to consider it is that for a given V_{ds} voltage, the current density in the waveguiding region will be greater in the double injection case than the single. Another benefit to double injection is that the electron and hole concentration over the entire channel are increased by an applied V_{ds} voltage. In the single injection device disucssed in the previous chapter, the hole injection is only into a thin region beneath the gate.

The disadvantage of the double injection structure is that an applied voltage on the insulated gate contact does not modulate the injection current I_{ds}. A positive gate voltage depletes holes under the gate region while attracting electrons. A negative gate voltage will decrease the electron concentration under the gate and increase the hole concentration.

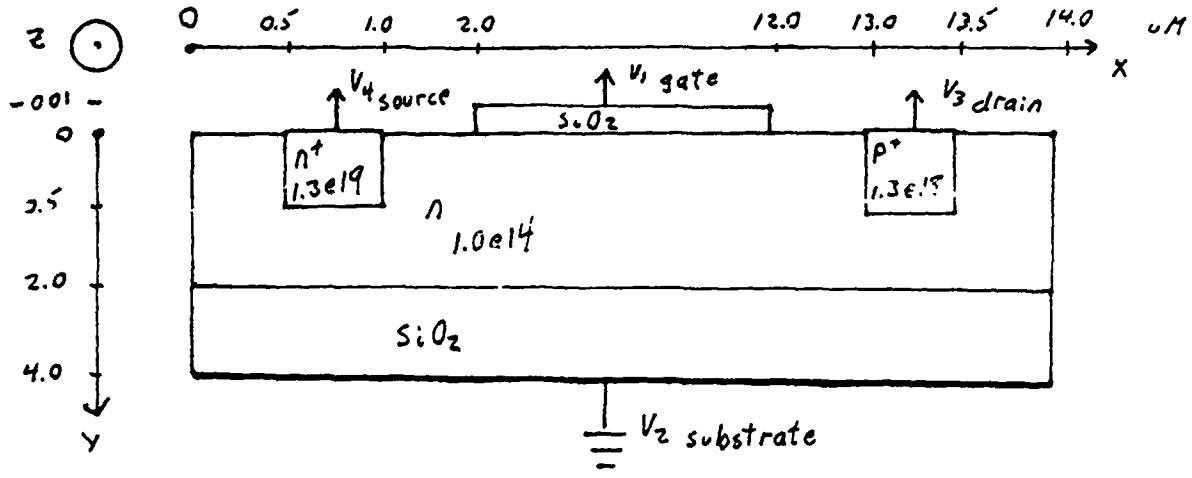
Using the PISCES IIB simulation program, two different simulations were run; one without including the effects of recombination (MDF120) and one including recombination effects (MDF140). The device geometry and doping profile are the same for MDF120 and MDF140 simulations. However, the model for simulation MDF140 includes Schockley-Reed-Hall (SRH) recombination effects to more accurately model bimolecular recombination. The MDF120 model is an ideal case and was used only to highlight the effects of recombination in the MDF140 model.

[1] Murray Lampert and Peter Mark, Current Injection in Solids, (Academic Press, New York, 1970), p. 208

5.1.1 Device Description

This device is similar to MDF220 except that this is a double injection MOSFET with a lower waveguide doping.

Figure 5: Double Injection MOSFET Waveguide for MDF120 and MDF140



Analyzing the two PISCES IIB simulations, MDF120 vs MDF140, it is possible to observe the results of bimolecular recombination in the waveguide structure.

5.1.2 Program Summary

Below is a summary of PISCES IIB programs run to characterize the device:

- MDF140 Define device and solve initial solution (zero bias).
- MDF141 Increase V_{ds} to 2.0 volts and create I-V curves for v1=0, v2=0, v3 from 0 to 2.0 volts and v4=0.
- MDF142 Increase gate (v1) from 0 to 5.0 volts with V_{ds} =1.0 volts (v3=1.0 v2=0 and v4=0).
- MDF143 Decrease gate (v1) from 0 to -5.0 volts with V_{ds} =1.0 volts (v3=1.0 v2=0 and v4=0).
- MDF124 find electron, hole, voltage and current density profile when v1=0, v2=0, v3=1.0 and v4=0.
- MDF145 find electron, hole, voltage and current density profile when v1=0.0, v2=0, v3=0.85 and v4=0.
- MDF146 Increase gate (v1) from 0 to 15 with V_{ds} = 0.85 volts (v2=0, v3=0.85 and v4=0).
- MDF147 Decrease gate (v1) from 0 to -25 with V_{ds} = 0.85 volts (v2=0, v3=0.85 and v4=0).
- MDF151 find electron, hole, voltage and current density profile when v1=5.0, v2=0, v3=0.85 and v4=0.

- MDF152 find electron, hole, voltage and current density profile when v1=-5.0, v2=0, v3=0.85 and v4=0.
- MDF153 find electron, hole, voltage and current density profile when v1=-15.0, v2=0, v3=0.85 and v4=0.
- MDF154 find electron, hole, voltage and current density profile when v1=15.0, v2=0, v3=0.85 and v4=0.
- MDF155 find electron, hole, voltage and current density profile when v1=-25.0, v2=0, v3=0.85 and v4=0.

The program name is used as the prefix to identify the three dimensional plots, contour plots and I-V curves. Note that the contour plots are not plotted to scale because the contour algorithm skews the dimensions.

5.2 Double Injection MOSFET: MDF140

Contour plots MDF140_THREED_N and MDF140_THREED_P are the equilibrium electron and hole contour plots of the double injection MOSFET device. In equilibrium, including recombination effects does not change the electron and hole distribution but the peak concentrations are lower compared to the idealized situation (MDF120). On the contour plots, the n^+ source region is on the left side of the device, near $x = 0$ uM and the p+ drain region is on the right, near $x = 14$ uM. PISCES IIB simulations MDF140 through MDF165 (all except MDF124) include Schockley-Reed-Hall recombination effects.

Table 5: Summary of Bias Voltage Conditions for MDF140 Double Injection

title	gate (v1)	substrate (v2)	drain (v3)	source (v4)
MDF140	0.0	0.0	0.0	0.0
MDF124	0.0	0.0	1.0	0.0
MDF145	0.0	0.0	0.85	0.0
MDF151	5.0	0.0	0.85	0.0
MDF152	-5.0	0.0	0.85	0.0
MDF153	-15.0	0.0	0.85	0.0
MDF154	15.0	0.0	0.85	0.0
MDF155	-25.0	0.0	0.85	0.0
MDF161	0.0	0.0	0.5	0.0
MDF164	10.0	0.0	0.5	0.0
MDF165	-10.0	0.0	0.5	0.0

5.3 MDF140 Drain-source Biasing

The difference between the single and double injection devices is that the single injection device is limited to only hole current but the double injection current is composed of both electrons and holes. Consequently, a double injection device will have a greater current flowing through it under similar biasing conditions than a single injection device.

Graph MDF141_IV is a plot of the drain to source current vs drain voltage (I_{ds} vs. V_{ds}) when the gate voltage is zero and recombination effects are included. Also the individual electron and hole currents vs. drain voltage were added. The curve shows that at lower voltages (less than 0.8 volts) the electron current is greater but above 0.8 volts, there is a greater contribution from the hole current.

Graph MDF_DOUBLE_COMPARE compares the I_{ds} current characteristics with and without recombination effects. If recombination is included, the current flow is greater until about 0.8 volts. At higher voltages, the MDF120 device (ideal case; recombination effects not included) will have greater current flow. This is expected because as the electron and hole densities increase, the rate of recombination would also increase. In this way the device would be limited by the current carriers lost through recombination in the device. Unfortunately, these excess carriers will be dissipated as excess heat.

5.4 MDF120 Analysis with $V_{ds} = 1.0$ volts.

In the single injection MOSFET, when the $V_{ds} = 1.0$ volts and the gate contact is zero volts, the peak current density was 10.1 A/cm^2. However, the current densities in the double injection are significantly higher. If V_{ds} = 1.0 volts, similar to single injection biasing, the waveguide electron and hole concentrations will both increase to about $1.0x18^{18}$ cm^{-3}; a very desirable result for phase modulation. However, the peak current density for this biasing configuration is over 50,000 A/cm^2 (MDF124_|JTOTAL|) which is unacceptably high.

5.5 MDF140 Analysis with $V_{ds} = 0.85$ volts.

If the effects of bimolecular recombination are considered, and V_{ds} = 0.85 volts, then the peak current density is 62 % greater than if recombination were not included. This would mean that the expected peak current density would be 2,733.5 A/cm^2 (MDF145_|Jtotal|). However, both the electron and the hole concentrations (MDF145_THREED_N, MDF145_THREED_P) are larger in the waveguiding region. This means that in a transverse device geometry, in order to get large electron and hole injections from the drain and source contacts, high peak current denisties in the waveguiding region are unavoidable. The expected concentrations would be a uniform change throughout the waveguide on the order of $1.0x10^{17}$ cm^{-3} for both electrons and holes.

5.5.1 Positive Biasing of MDF140 Gate Electrode with V_{ds} = 0.85 volts

If the gate contact is biased positive to +5.0 volts with V_{ds} = 0.85 volts, the total I_{ds} current does not change. Contour plot MDF151_THREED_P (B) is a log plot of the concentrations from 16.5 to 17.0 cm^{-3} which shows a slight depletion of holes beneath the gate region. There is also a corresponding electron contour plot that shows the increased electron concentration in the same region.

A potential problem with this type of biasing is that the electron concentration underneath the gate is increased $2.0x10^{17}$ cm^{-3} while the hole concentration is decreased $0.3x10^{17}$ cm^{-3}. A quick calculation at 1.3 uM would predict a net index change of

$$\Delta n = -6.2x10^{-22}\Delta N_e - 6.0x10^{-18}\Delta N_h^{0.8} \quad 5.1$$

$$\Delta n = -6.2x10^{-5} + 9.12x10^{-5} \quad 5.2$$

$$\Delta n = 2.92x10^{-5} \quad 5.3$$

The electron and hole changes are not adding together to form larger index changes because holes are depleted while electrons are attracted. The index change profiles, MDF151_INDEX_13 and MDF151_INDEX_155 show that the predicted Δn is positive throughout the MOSFET indicating that the overall effect in the waveguiding region is that holes are depleted.

The double injection effects may cancel out instead of adding together under certain conditions. The peak current density has not changed from the gate = 0 volt case (mdf151_|Jtotal|).

If the applied gate voltage is increased to 15.0 volts, there is an increase in the electron concentration under the gate (MDF154_THREED_N). There is also a slight change in the hole concentration (MDF154_THREED_P) but the largest hole concentration change appears outside the waveguiding region. Comparing index changes plots at 1.55 uM wavelengths, MDF151_INDEX_155 and MDF154_INDEX_155, the only area where a +15.0 gate voltage is an improvement from +5.0 is in the depletion width beneath the gate.

Based on the two simulations with a positive gate voltage, MDF151 and MDF154, there is going to be a maximum voltage that will deplete holes without attracting a large concentration of electrons. Since depletion will increase the index of refraction and accumulation decreases the refractive index, the change in electron and hole concentrations could offset each other instead of adding.

5.5.2 Negative Biasing of MDF140 Gate Electrode with V_{ds} = 0.85 volts

A negative 5 volts bias applied to the gate contact produces the opposite effect of a positive 5 volts. There is a slight depletion of electrons under the gate (MDF152_THREED_N) while there is a corresponding increase in the hole concentration (MDF152_THREED_P) in the same region. The current density profile (MDF152_|Jtotal|) is not noticeably different.

The index change profiles have two distinct regions; one underneath the gate where Δn is negative and the bulk of the waveguide where Δn is positive. Contour plots MDF152_INDEX_155 and MDF152_INDEX_13 are plots of the predicted index change at 1.55 uM and 1.3 uM wavelengths.

If the gate voltage is decreased to -15.0 volts, then the maximum index change at 1.55 uM will be -0.00127. This is a third order effect produced when a sufficient concentration of holes is attracted under the gate to offset the depletion of electrons. The peak current density has increased to 2,856 A/cm2 (MDF153_|Jtotal|). Note also that the current density under the gate region has increased.

When the gate voltage is decreased to -25.0 volts, then the predicted index change at 1.55 uM will be -0.00203. However, the current density is also greater (MDF155_|Jtotal|).

As the gate voltage becomes more negative, the shift in electron and hole concentrations create two different areas within the waveguiding structure. The index change of the bulk of the structure will get smaller (but still positive) due to holes being depleted to the gate region while beneath the gate region, the hole concentration is becoming large enough to offset the depletion of electrons. So under the gate region, the index change will get larger (and more negative) with a more negative gate voltage.

5.6 MDF140 Analysis with $V_{ds} = 0.5$

If the V_{ds} voltage is 0.5 volts, then the current density is very low because V_{ds} is less than the built-in voltage. In the waveguiding region, the electron concentration increases to $1.0x10^{15}$ (MDF161_THREED_N) while the holes are are slightly depleted (MDF161_THREED_P).

5.6.1 Positive Biasing of MDF140 Gate with $V_{ds} = 0.5$

A positive gate voltage will attract a large concentration of electrons underneath the gate (MDF164_THREED_N) without depleting holes. Although, the hole concentration is decreased, it is not large enough to affect the refractive index.

5.6.2 Negative Biasing of MDF140 Gate with $V_{ds} = 0.5$

A negative gate voltage will attract a large concentration of holes beneath the gate (MDF165_THREED_P). The electrons are only depleted from top 0.1 uM waveguide region. The middle of the waveguide exhibits an increased electron concentration (MDF165_THREED_N) because the electrons depleted by the negative gate voltage increase the electron concentration in the middle of the waveguide.

5.7 Results and Discussion

Based on the PISCES IIB MOSFET simulations, double injection is a good technique to simultaneously inject electrons and holes into the waveguiding region. Further, unlike a single injection situation, the MOSFET is uniformly injected with a higher electron and hole concentration. In a single injection MOSFET, most of the current flow (therefore, the injected holes as well), is directly underneath the contact.

Unlike the single injection cases though, the applied gate voltage does not modulate the I_{ds} current. A positive gate voltage induces a slight shift from holes to electron I_{ds} current while a negative gate voltage causes a slight change from hole to electron I_{ds} current. The overall I_{ds} current is not significantly changed.

An applied gate voltage will not significantly alter the electron and hole distributions of a double injection MOSFET beyond the depletion width. The maximum depletion width, W_m, is calculated from the concentration and, since V_{ds} = 0.85 volts will increase the electron and hole concentration to about $1x10^{17}cm^{-3}$, the maximum depletion width is [2]

$$W_m = \sqrt{\frac{4\epsilon_s K_b T \ln(\frac{N_a}{n_i})}{q^2 N_a}} \quad 5.4$$

where

ϵ_s is the permittivity F/cm,

N_a is the doping concentration,

and n_i is the intrinsic concentration.

[2] S. M. Sze, Semiconductor Devices Physics and Technology, (Wiley & Sons, New York, 1985), p. 191.

Solving for W_m

$$W_m = \sqrt{\frac{4 * 11.9 * 8.85x10^{-14} * 0.0258 \ln(\frac{1x10^{17}}{1.45x10^{10}})}{1.602x10^{-19}1x10^{17}}} \quad 5.5$$

$$W_m = 0.103\ uM \quad 5.6$$

The index changes are limited to the 0.1 uM layer underneath the gate contact.

For an insulated gate waveguide structure, the effective index change is not significantly altered when the waveguide is 2.0 uM thick. However, if only consider the 0.1 uM under the gate region is considered, then the effective index changes vs gate voltage will be significant.

As shown in earlier work, double injection creates a very rapid increase in channel current with increasing V_{ds}. [3] If a more realistic recombination model is included in the simulation, then the predicted current densities will be very high (2,500 to 3,000 A/cm^2) for this geometry. Therefore, the first conclusion is that, for the transverse silicon waveguide, the current density is going to be the limiting factor. This is the same conclusion predicted by Friedman, Soref and Lorenzo in their work. [4]

Below is a table summarizing the predicted peak currents vs gate voltage for the MDF140. Also in the table is the predicted peak current densities for the ideal case, MDF120.

Table 6: Summary of Peak Current Density vs Gate Voltage with V_{ds} = 0.85 volts

Gate voltage (v)	peak J A/cm^2 includes recombination	peak J A/cm^2 without recombination
0	2733.5	1685.7
5.0	2763.5	1684.8
15.0	2860.2	1691.6
-5.0	2758.1	1693.3
-15.0	2856.0	1718.7
-25.0	2953.2	

However, this device has potential as a silicon optical waveguide modulator. With a large negative bias applied to the gate contact with V_{ds} = 0.85 volts, a maximum Δn of greater than 0.001 is predicted. Below is a table summarizing the maximum index change expected vs gate voltage. It should be kept in mind that the peak negative index changes are directly under the gate while the positive changes are in the middle of the device.

[1] M. Hack, M. Shur and W. Czubatyj, "Double-Injection Field- Effect Transistor: A new Type of Solid-State Device", Applied Physics Letters, May 19, 1986, 1387.

[4] L. Friedman, R. Soref and J. Lorenzo, "Silicon Double- Injection Electro-Optic Modulators with Junction Control", Journal of Applied Physics, March 31, 1988, p. 1837.

Table 7: Maximum Index changes vs Gate Voltage for MDF140 Waveguide

title	V_{ds}	initial gate (v)	final gate (v)	max $\Delta n(1.3)$	max $\Delta n(1.55)$
MDF151	0.85	0	5.0	5.90-54	8.40e-5
MDF152	0.85	0	-5.0	-2.98e-4	-4.23e-4
MDF153	0.85	0	-15.0	-8.96e-4	-1.27-3
MDF154	0.85	0	15.0	-1.73e-4	-2.46e-4
MDF155	0.85	0	-25.0	-1.43e-3	-2.03e-3
MDF164	0.50	0	10.0	-2.59e-4	-3.66e-4
MDF165	0.50	0	-10.0	-7.13e-4	-1.01e-3

In order to compare biasing scheme of the waveguides, the effective index change, Δn_{eff} was also calculated. The integral limits are the defined boundaries of the waveguide in the x and y direction. In this case, the x limits are the edges of the source and drain regions at 1.0 uM and 13.0 uM.

Below is a summary table listing the effective index of the entire 12 uM wide (distance from drain to source) by 2.0 uM deep waveguiding region.

Table 8: Summary of Effective Refractive Index Changes vs Gate Voltage for a 24 uM^2 Guide

title	V_{ds}	initial gate (v)	final gate (v)	$\Delta n_{eff}(1.3)$	$\Delta n_{eff}(1.55)$
MDF151	0.85	0.0	5.0	-4.52e-6	-6.41e-6
MDF152	0.85	0.0	-5.0	-3.78e-6	-5.36e-6
MDF153	0.85	0.0	-15.0	-1.29e-5	-1.83e-5
MDF154	0.85	0.0	15.0	-1.47e-5	-2.09e-5
MDF155	0.85	0.0	-25.0	-1.96e-5	-2.78e-5
MDF164	0.50	0.0	10.0	-7.05e-8	-9.98e-8
MDF165	0.50	0.0	-10.0	-7.76e-7	-1.62e-6

For an insulated gate waveguide structure, the effective index change is not significantly altered when the waveguide is 2.0 uM thick. However, if only the 0.1 uM under the gate region is considered then the effective index changes vs gate voltage will be significant. This data is summarized in the next table.

Table 9: Summary of Effective Refractive Index Changes vs Gate Voltage for a 1.2 uM² Guide

title	V_{ds}	initial gate (v)	final gate (v)	$\Delta n_{eff}(1.3)$	$\Delta n_{eff}(1.55)$
MDF151	0.85	0.0	5.0	2.70e-5	3.91e-5
MDF152	0.85	0.0	-5.0	-1.43e-4	-2.06e-4
MDF153	0.85	0.0	-15.0	-4.24e-4	-6.12e-4
MDF154	0.85	0.0	15.0	-6.54e-5	-9.43e-5
MDF155	0.85	0.0	-25.0	-6.76e-4	-9.77e-4
MDF164	0.50	0.0	10.0	-1.12e-4	-1.62e-4
MDF165	0.50	0.0	-10.0	-3.39e-4	-4.90e-4

Based on the double injection MOSFET simulations, one can draw the following conclusions.

- A double injection MOSFET will have significantly higher electron and hole concentrations than a single inject MOSFET.
- Unlike the longitudinal DIFET, An insulated gate structure will not have the electron and hole concentrations in the waveguide region pinched off due to an applied V_{ds} voltage.
- For the ideal case, a negative gate voltage which attracts holes while depleting electrons will induce a larger index change than a positive gate voltage.
- The V_{ds} voltage will inject carriers throughout the device and create a uniform index change. This will produce a symmetrical waveguide.
- The V_g voltage only depletes the calculated W_m region.
- Peak refractive index changes will approach 1e-3 with sufficient negative gate voltage.

CHAPTER 6

DOUBLE INJECTION LONGITUDINAL FET ELECTRO-OPTIC MODULATOR

6.1 Introduction

A large portion of this study was devoted to analyzing the double injection long channel junction field effect transistor as a possible design for a silicon waveguide. From a one dimensional analysis, this type of structure offered the potential of injecting large electron and hole concentrations into the conduction channel without the high current densities expected in a transverse design. [1]

The one dimensional model made certain assumptions to simplify the longitudinal analysis such as assuming one-dimensional planar current flow, neglecting diffusion currents and simplifying recombination effects. [2] Using the PISCES IIB semiconductor simulation program, a more complete two dimensional model of the DIFET was designed and characterized. PISCES IIB solves Poisson's equation and the continuity equations for electrons and holes to obtain the voltage potential, electron, hole and current density distributions for a given set of doping and biasing conditions.

Phase modulation of an optical signal is similar for transverse and longitudinal waveguide geometries. The silicon waveguide region is bound by lower refractive index material such as heavily doped silicon or oxide. Free carriers are injected into the waveguide region to alter the refractive index and change the propagation of optical signals through the waveguide. If the waveguide has a heavily doped gate region as a third contact, then a negative gate voltage will remove the free carriers in the conduction channel.

In the longitudinal design, light travels along the x axis, parallel to the direction of current flow from the anode-cathode contacts, and is confined on the y axis by a heavily doped gate and insulated substrate region. For a DIFET, the optical waveguiding region is the conduction channel. This is analogous to a slab waveguide [3] and an optical signal is phased modulated in proportion to the length of the waveguide.

However, the longitudinal design is based on the assumption that an applied voltage on the anode and cathode contacts will inject large concentrations of carriers to induce refractive index changes. Based on two dimensional PISCES IIB device simulations, the high electron and hole densities will not be injected into the semiconductor plasma under the gate contact region. Instead, there are limitations on the maximum bias voltages that can be applied to the anode and cathode contacts before the heavily doped gate or conduction channel is affected.

[1] Lionel Friedman, Richard Soref and Joseph Lorenzo, "Silicon Double-Injection Electro-Optic Modulators with Junction Gate Control", Journal of Applied Physics, March 31, 1988, p. 1831.

[2] Lionel Friedman, "Proposal to the Air Force Office of Scientific Research", Research Initiation Program, December, 1987, p. 1.

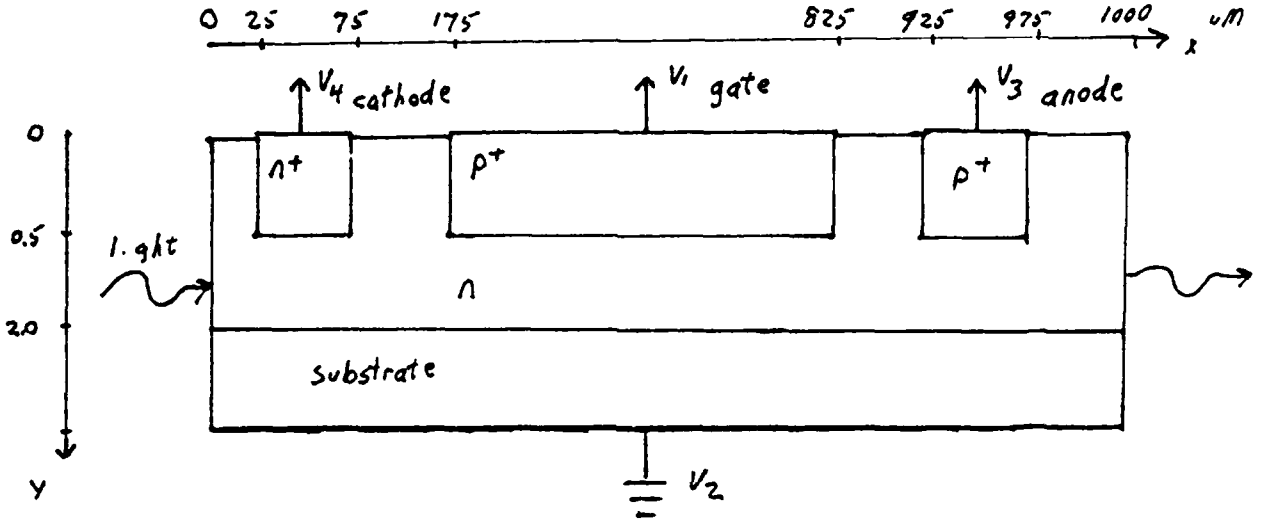
[3] Dietrich Marcuse, Theory of Dielectric Optical Waveguides, (Academic Press, New York, 1974), p. 13.

Electron injection from the cathode contact is limited because the cathode and gate region form a built-in p-i-n junction. When the cathode contact is biased below -0.7 volts, the electron concentration is increased *between* the two regions but not under the gate region.

Hole injection from the anode contact is limited because a bias voltage on the anode high enough to inject sufficient concentrations of holes under the gate region pinches off the conduction channel. The result is a much smaller waveguiding region than expected.

6.1.1 Longitudinal Device Description

Figure 6: Longitudinal Double Injection FET with Junction Gate Control



The longitudinal DIFET is 1000 uM long and 4 to 5 uM thick with a 650 uM gate contact centered on the top of the device. The anode and cathode regions are 50 uM long by 0.5 uM deep at each end. An applied bias on the anode and cathode contacts will inject electrons and holes into the device and the waveguiding region is the conduction channel through the middle of the structure which is doped $1x10^{16}$ n type.

The four electrodes are neutral ohmic contacts. The early test cases did not include recombination effects in the model, but the final model, DIF310, included Schockley-Reed-Hall recombination effects.

The gate region was always heavily doped p+ type similar to a junction FET where the applied negative gate voltage controls the effective cross-sectional area of the conducting n-type channel. [4] In a double injection simulation, the cathode region is doped n+ type and the anode is p+ type. A potential problem with this design is that a negative voltage applied to the n+ cathode, will attract holes not only from the p+ anode, but also from the gate. If the cathode voltage is large enough, then most of the current is sourced from the gate, not the anode. The result is an increase in hole concentration between the cathode and gate but not under the gate.

[4] Ben Streetman, Solid State Electronics Devices, (Prentice-Hall, Englewood Cliffs, New Jersey, 1980), p 286.

Four test cases were simulated with different doping concentrations in the gate and anode regions. The cathode voltage in the PISCES IIB simulation was decreased from 0 to -1.0 volts in -0.1 volt steps and the current-voltage characteristics of the gate and anode contact were graphed on a semilog scale.

Table 10: Summary of Longitudinal I-V Graphs and Doping Concentrations for Gate Designs

graph	cathode doping	gate doping	anode doping
DIF211_IV	1.3e19 n+	1.3e19 p+	1.3e19 p+
DIF261_IV	1.3e19 n+	1.0e16 p+	1.3e19 p+
DIF271_IV	1.3e19 n+	1.3e19 p+	1.0e17 p+
DIF401_IV	1.3e19 n+	1.3e19 p+	1.3e19 n+

Graph DIF211_IV is a graph of the log of the current from the anode and the gate vs. negative cathode voltage. From the graph, the gate sources more current than anode when they are both heavily doped p+ type. Furthermore, below -0.5 volts, the anode current begins to level off while the gate continues to increase significantly. From graph DIF211_IV, one can conclude that equally doping the anode and gate regions will not produce a large hole injection when the cathode is reverse biased.

Next, the gate doping was decreased from $1.3x10^{19}$ p+ to $1x10^{16}$ p+. The effect was that all the current was sourced from the gate contact and none from the anode (graph DIF261_IV).

Finally, the anode doping was decreased to $1.0x10^{17}$ while the gate doping remained $1.3x10^{19}$. The current-voltage curves, DIF271_IV, of the gate and anode indicate that from 0 to -0.7 volts, the anode will inject more holes into the plasma than the gate. But below -0.7 volts the gate current is very similar to the first case, DIF211_IV.

From these curves, one can conclude that the cathode doping must be less than the gate doping or a negative voltage on the n+ anode region will source current from the gate not the cathode. If the gate region is the main source of holes, then holes are injected between the anode and gate but not in the waveguide region. A bias voltage on the p+ anode contact will inject holes into the waveguiding region but it also causes the conduction channel to pinch off.

The substrate region must be either heavily doped silicon or an oxide layer to confine the light. The substrate material must be selected carefully because the waveguide is so thin. In the longitudinal design the anode and cathode are 100 uM from the gate but only 1.5 uM from the substrate.

Table 11: Summary of Longitudinal Programs to find Optimal Substrate material

title	gate	waveguide	substrate
DIF160	1.3e19 p+	1.0e16 n	1.3e19 n+
DIF170	1.3e19 p+	1.0e16 n	oxide
DIF180	1.3e19 p+	1.0e16 n	1.3e19 p+

If the substrate layer is heavily doped n+ silicon and at zero potential (ground), then a negative cathode voltage will inject holes between the n+ cathode and the n+ substrate, not into the waveguide. Biasing the anode positive, gives similar results. Holes from the anode are swept out by the gate and substrate; not the through the device to the cathode contact.

A heavily doped p+ substrate would be an ideal good choice for the substrate because a device with a p+ gate and p+ substrate would be able to pinch off the channel from the top and bottom simultaneously. This would make the waveguiding region symmetrical since it would be bounded, top and bottom, by identical confinement layers. Furthermore, since the gate and substrate would be reverse biased together, the waveguiding region will remain symmetrical until pinched off completely.

However, when the n+ cathode is biased negatively, both the gate and substrate source more current than the anode. This is shown on graph DIF181_IV. Forward biasing the anode contact (shown on graph DIF182_IV) does not provide sufficient electron or hole injection either because most of the injected electrons are swept away by the p+ substrate, not the gate.

In a longitudinal design, the waveguide is very thin. and the substrate is going to be near the doping regions. Therefore, oxide would make the best substrate material since an insulated substrate would not sweep out any electrons or holes. The oxide also provides a strong confinement layer since it has such a low index of refraction.

6.1.2 Early Test Cases

DIFETs DIF160 through DIF250 are early test cases that will not be discussed in great detail. These simulations did not include recombination effects but were useful in exposing problems with the *DIFET* geometry.
Based on the first test cases of simulated longitudinal double injection FETs one can draw the following conclusions:

- The current density is very low compared to transverse waveguide devices.
- If the predicted index changes are less than 0.001 in the waveguiding region, then the device will have to be very long to create a π radian phase change.
- If the gate and cathode are doped equally, then a negative voltage on the n+ anode contact will source holes from the gate, not the cathode.
- A negative cathode voltage injects electrons into the waveguide but, regardless of the doping profile, the p-i-n junction from the cathode-gate region will limit the negative bias voltage that can be applied to the cathode. Consequently, the electron injection capabilites of the longitudinal DIFET will be limited.
- A heavily doped silicon substrate may cause problems when the anode or cathode contacts are biased because the substrate contact will sweep away holes.
- Increasing the anode voltage beyond 4.0 volts, causes the conduction channel to start to contract near the anode contact.
- A negative gate voltage will sweep out all electrons and holes from the conduction channel which is the confined region between the gate and substrate.

6.2 Phase Change vs. Device Length

Electro-optic materials can be used as optical phase modulators by applying an electric field over the length of the waveguide. The change in phase can be expressed as [5]

$$\Delta\phi = \frac{\pi}{\lambda} n_{eff}^3 r E \quad 6.1$$

where

λ is the wavelength,

n_{eff}^3 is the effective refractive index,

r is the electro-optic coefficient (cm/V) in the direction of propagation,

and E is the applied electric field.

However, this equation only applies to birefringent crystals. A more general definition of the phase shift is [6]

$$\Delta\phi = \Delta\beta L \quad 6.2$$

where

β is the propagation constant, $k_o n$, wavenumber times the refractive index,

and L is the length of the waveguide.

Since [7]

$$\Delta\beta = \frac{2\pi}{\lambda}\Delta n \quad 6.3$$

the phase shift can be expressed as

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta n L \quad 6.4$$

where

Δn is the change in the refractive index of the waveguide.

If one assumes an index change of 0.001 in the waveguide region at 1.3 uM, then for a phase change of π radians, the length would need to be 650 uM.

Obviously, the length of the device could be changed, but 650 uM is sufficiently long to neglect any short channel effects in a semiconductor waveguide design.

---

[5] Amnon Yariv and Pochi Yeh, Optical Waves in Crystals, (Wiley, New York, 1984) p. 283.

[6] Leon McCaughn, "Advanced Guided-Wave Integrated Optic Devices", short course at SPIE, Boston, MA, September 6, 1988, p. 22 class notes

[7] Y. Suematsu and K. Ichiiga, Introduction to Optical Fiber Communication, (Wiley & Sons, New York, 1982), p. 24.

6.3 Analysis Method

1. Design longitudinal DIFFT using PISCES IIB simulation programs.
2. Analyze the electrical characteristics of the semiconductor to find the maximum increase in the free carrier concentration in the waveguide region.
3. Gather data on the electron, hole and current density distributions for different biasing conditions.
4. Calculate the change in the refractive index at each mesh point.
5. Calculate the total phase shift of the waveguide vs. length.

CHAPTER 7

DOUBLE INJECTION LONGITUDINAL SILICON WAVEGUIDE WITH AN OXIDE SUBSTRATE

7.1 Introduction

The longitudinal double injection field effect transistor (DIFET) as an optical phase modulator was first proposed in the paper by Friedman, Soref and Lorenzo. [1] This work further develops their ideas on DIFETs by analyzing a more complete model of the double injection semiconductor. The electrical characteristics were modeled using the semiconductor simulation program PISCES IIB and the electron and hole data was used to calculate the optical properties of the device.

Two different models were analyzed; one without recombination effects and one including Schockley-Reed-Hall recombination. This was done so that the effects of bimolecular recombination could be observed. Model DIF270 is evaluated without including recombination effects while model DIF310 includes Schockley-Reed-Hall recombination. The same set of programs were run to characterize the two models so only the differences between DIF270 and DIF310 will be mentioned.

The model was designed with the following conditions:

- The doping regions are 0.5 uM deep abrupt junctions.
- The gate is heavily doped p+.
- The cathode is heavily doped n+ and the anode is lightly doped p type.
- The incoming optical signal will only be phase modulated between the gate and substrate.
- The substrate is insulated to prevent the electrons and holes from being swept out.

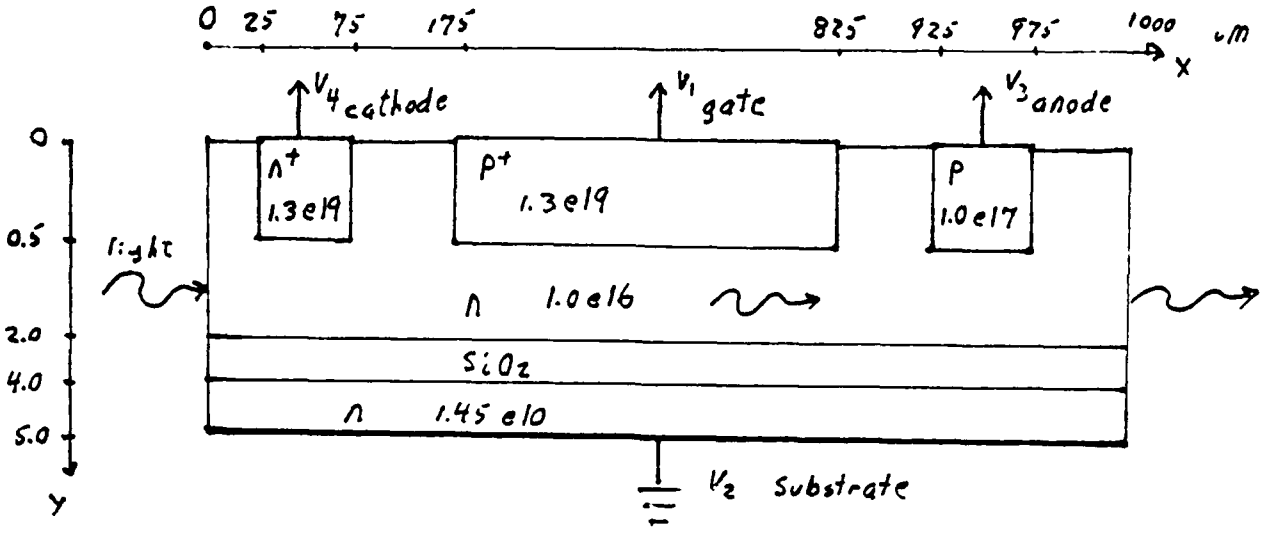
Although there are increases in the electron and hole concentrations due to double injection, the changes need to be kept in perspective. If the concentration change is $< 1.0x10^{16}$ cm^{-3} then the refractive index of the silicon waveguide will not differ significantly from that of intrinsic silicon. [2] Although the concentration changes are positive developments, they are not large enough to design a phase modulator.

---

[1] Lionel Friedman, Richard Soref and Joseph Lorenzo, "Silicon Double-injection Electro-optic Modulators with Junction Gate Control", Journal of Applied Physics, March 15, 1988, p. 1831-9.

[2] Richard Soref and Joseph Lorenzo, "All-Silicon Active and Passive Guided-Wave Components for 1.3 and 1.6 uM wavelengths", IEEE Journal of Quantum Electronics, Vol. QE-22, June, 1986, p. 874.

Figure 7: Longitudinal DIFET Model for DIF270 and DIF310



7.1.1 Device Description

The four terminal simulated structure was 1000 uM long and 5.0 uM deep. The substrate was intrinsic silicon with a layer of oxide between the substrate and the conduction channel. The waveguiding region was doped $1x10^{16}$ cm^{-3} n type and extended along the entire device, 1.5 uM thick. The gate region was a heavily doped $1.3x10^{19}$ cm^{-3} p+ region, 650 uM long and 0.5 uM deep centered in the device. The anode region was a 50 uM long 0.5 deep doped p $1x10^{16}$ cm^{-3} region at the end of the device while the cathode was the same dimensions but heavily doped n+ $1.3x10^{19}$ cm^{-3} at the beginning of the device.

7.1.2 Program Summary

The following PISCES IIB programs were run to characterize the device:

- DIF310 Define device and solve initial solution with recombination effects.
- DIF311 Decrease voltage on anode V_4 with $V_1=0$, $V_2=0$ and $V_3=0$.
- DIF312 Find electron, hole and current density distributions for $V_1=0.0$, $V_2=0.0$, $V_3=0.0$ and $V_4=-0.7$.
- DIF313 Find electron, hole and current density distributions for $V_1=0.0$, $V_2=0.0$, $V_3=0.0$ and $V_4=-0.5$.
- DIF314 Increase voltage on cathode V_3 with $V_1=0$, $V_2=0$ and $V_4=-0.7$.
- DIF315 Find electron, hole and current density distributions for $V_1=0.0$, $V_2=0.0$, $V_3=0.7$ and $V_4=-0.7$.
- DIF316 Decrease gate voltage to -17.5 with $V_2=0.0$, $V_3=0.0$ $V_4=-0.7$.
- DIF317 Decrease gate voltage to -17.5 with $V_2=0.0$, $V_3=0.0$ $V_4=-0.5$.
- DIF318 Decrease gate voltage to -17.5 with $V_2=0.0$, $V_3=0.7$ $V_4=-0.7$.
- DIF321 Find electron, hole and current density distributions for $V_1=-17.5$, $V_2=0.0$, $V_3=0.0$, and $V_4=-0.7$.

- DIF322 Find electron, hole and current density distributions for $V_1 = -17.5$, $V_2 = 0.0$, $V_3 = 0.0$, and $V_4 = -0.5$.
- DIF323 Find electron, hole and current density distributions for $V_1 = -17.5$, $V_2 = 0.0$, $V_3 = 0.7$, and $V_4 = -0.7$.
- DIF331 Increase voltage on cathode V_3 with $V_1 = 0$, $V_2 = 0$ and $V_4 = 0$.
- DIF332 Find electron, hole and current density distributions for $V_1 = 0.0$, $V_2 = 0.0$, $V_3 = 0.7$ and $V_4 = 0$.
- DIF333 Find electron, hole and current density distributions for $V_1 = 0.0$, $V_2 = 0.0$, $V_3 = 4.0$ and $V_4 = 0$.

7.2 Double Injection DIFET: DIF310

Contour plots of the electron (DIF310_THREED_N) and hole (DIF310_THREED_P) concentrations show the doping regions of the device. The n+ type cathode is near x = 0 and the p+ type anode is near x = 1000 uM. The line through middle of the electron plot is the waveguide-oxide boundary.

The potential at the anode contact is

$$V_{contact} = \frac{k_b T}{q} \ln\left(\frac{Na}{ni}\right) \quad 7.1$$

$$V_3 = (0.0258) \ln\left(\frac{2.2x10^{18}}{1.45x10^{10}}\right) \quad 7.2$$

$$V_3 = 0.4860eV \quad 7.3$$

and the potential at the cathode contact is

$$V_4 = -(0.0258) \ln\left(\frac{1.0x10^{16}}{1.45x10^{10}}\right) \quad 7.4$$

$$V_4 = -0.3469eV \quad 7.5$$

7.3 Anode and Cathode Bias Voltages

Below is a summary of the bias conditions simulated on the DIF310 longitudinal DIFET.

Table 12: Summary of Bias voltages for DIF310

title	gate (v1)	substrate (v2)	anode (v3)	cathode (v4)
DIF310	0.0	0.0	0.0	0.0
DIF312	0.0	0.0	0.0	-0.7
DIF313	0.0	0.0	0.0	-0.5
DIF315	0.0	0.0	0.7	-0.7
DIF332	0.0	0.0	0.7	0.0
DIF333	0.0	0.0	4.0	0.0

Reverse biasing the n+ type cathode contact will attract holes into the waveguide from the anode and repel electrons from the cathode region. However, the p+ gate and n+ cathode portion of the device form a p-i-n junction and there is a lower limit of about -0.5 volts on the cathode (v4) voltage. Graph DIF311_IV is a plot of the increase in gate and anode current vs negative gate voltage, v_1, (all other contacts at zero volts). The curves show that from 0 to -0.5 volts the anode current is roughly equal to the gate current. However, below -0.5 volts the current sourced from the gate region contact is greater than the current from the anode contact.

Forward biasing the p+ anode contact attracts electrons into the waveguide from the cathode and repels holes from the anode region. The I_{ca} current increases with cathode voltage and the gate current (holes swept out by the gate contact) is approximately 1% of the total I_{ca} current. Curves of the current vs. cathode voltage are found on graph DIF331_IV.

7.3.1 Reverse Biasing the Cathode Contact

Contour plot DIF312_THREED_P shows the increased hole concentrations due to a negative bias on the cathode. The holes are increased from $2.0x10^6$ to about $1.0x10^{13}$ cm^{-3} in the waveguiding region. There is an even larger increase from the gate as holes are injected ($1.0x10^{14}$) between the gate and cathode contact. Electrons (DIF312_THREED_N) are repelled from the cathode and there is a slight increase in electron concentration under the cathode.

The estimated peak current density (DIF312_|JTOTAl|) is only 60 A/cm^2 between the gate and cathode regions. Decreasing the cathode voltage beyond -0.7 volts will not inject a significant concentration of holes into the waveguide from the anode because the increased hole concentration will be offset by a large increase in gate current and injected holes outside the gate region.

A negative 0.5 volts applied to the cathode contact increases the hole concentration (DIF313_THREED_p) to $1.0x10^{12}$ uniformly across the waveguiding region. At -0.5 volts, There is a minimal contribution of holes from the gate region. The electron concentration (DIF313_THREED_N) is evenly distributed throughout the waveguide region.

7.3.2 Forward Biasing the Anode Contact

If the anode contact is forward biased to 0.7 volts, then the electron concentration (DIF332_THREED_N) is not significantly different from equilibrium. The hole concentration (DIF332_THREED_P) is increased slightly but only near the anode region.

A 4.0 volt bias voltage on the anode increases the hole concentration in the anode region but does not increase the hole concentration (DIF333_THREED_P) in the waveguide above $1x10^{12}$ cm^{-3}. The electron concentration (DIF333_THREED_N) between the gate and anode is increased but the electron concentration in the waveguide is beginning to get pinched off in the conduction channel.

7.3.3 Forward Biasing Anode 0.7 Volts Reverse Biasing the Cathode to -0.7 Volts

The maximum cathode voltage that can be applied without a significant contribution from the gate contact is -0.7 volts. Then, increasing the anode contact to +0.7 volts will increase the electron concentration inside the waveguide. The electron concentration (DIF315_THREED_N) plot shows that the electrons are injected from the n+ negatively biased cathode at x=0. However, under the gate, the concentration is still slightly below $1.0x10^{16}$ cm^{-3}.

When there is a negative bias on the cathode contact, the problem with increasing the anode voltage to inject holes is that it creates an area void of holes (DIF315_THREED_P) under the gate region. Compared to only the cathode biased (DIF312), there is still a contribution from the gate and a slight increase in the hole concentration near the anode.

The maximum current density (DIF315_ | JTOTAL |) is 67 A/cm^2 but the current is more evenly distributed throughout the waveguide compared to the DIF312 test cases.

7.4 Decreasing the Gate Voltage to Pinchoff

The pinchoff voltage is defined as [3] .

$$V_p = \frac{qa^2 N_d}{2\epsilon_s} \quad 7.6$$

where

q is the electronic charge,

a is the radius of the conduction channel,

N_d is the donor concentration,

and ϵ_s is the permittivity.

For the DIFET, the calculated pinchoff voltage is

$$V_p = \frac{1.602x10^{-19}(0.75x10^{-04})^2 4e16}{2 * 11.9 * 8.85x10^{-14}} \quad 7.7$$

$$V_p = 17.1 \ volts \quad 7.8$$

Decreasing the gate voltage will pinchoff the channel under the gate and any electrons or holes injected from the anode and cathode will be swept out. Reverse biasing the gate to -17.50 volts will deplete the channel regardless of the anode-cathode biasing voltages.

Since the pinchoff voltage is proportional to the conduction channel concentration, increasing N_d will require a proportional increase in V_p to pinch off the injected charge.

Table 13: Summary of Bias Gate Pinchoff voltages for DIF310

title	gate (v1)	substrate (v2)	anode (v3)	cathode (v4)
DIF310	0.0	0.0	0.0	0.0
DIF321	-17.5	0.0	0.0	-0.7
DIF322	-17.5	0.0	0.0	-0.5
DIF323	-17.5	0.0	0.7	-0.7

Contour plots DIF321_THREED_N and DIF321_THREED_P are typical of the waveguide when the gate is reverse biased. The waveguide region under the gate is depleted of electrons ($< 1.0x10^{10}$) and holes ($< 1.0x10^{10}$) while the region outside the gate is unchanged.

[3] Streetman, p. 290

7.5 Index of Refraction Changes

The change in index of refraction is calculated from the change in the electron and hole distributions. In silicon, changes in the real component of the refractive index, Δn, can be calculated using data from Richard Soref. [4] .

$$\Delta n_e(1.3) = -6.2x10^{-22}\Delta N_e \quad 7.9$$

$$\Delta n_h(1.3) = -6.0x10^{-18}(\Delta N_h)^{0.8} \quad 7.10$$

At 1.55 uM wavelength, the predicted refraction effect is

$$\Delta n_e(1.55) = -8.8x10^{-22}\Delta N_e \quad 7.11$$

$$\Delta n_h(1.55) = -8.5x10^{-18}(\Delta N_h)^{0.8} \quad 7.12$$

There is a linear relationship between the index of refraction and the electron concentration but a nonlinear relationship between Δn and the hole concentration. Since the index change constants are negative, electron or hole *injection* will *decrease* the refractive index while *depletion* will *increase* the refractive index.

Finally, the refractive index changes due to the change in electron concentration and change in hole concentration at each xy mesh point are added together to form the total refractive index change profile.

$$\Delta n = \Delta n_e + \Delta n_h \quad 7.13$$

Contour plots DIF321_THREED_13 and DIF321_THREED_155 are typical of the index change profiles expected from this geometry. Although, there is a large 10^{-4} increase in index change, it is in the gate region not the waveguiding region. The refractive index change in the conduction channel is on the order of $1.0x10^{-6}$, much too small for phase modulation.

Table 14: Summary of Maximum Index Change in Gate Region when V_1 is changed from 0 to -17.5 volts

title	anode (v3)	cathode (v4)	$\Delta n(1.3)$	$\Delta n(1.55)$
DIF321	0.0	-0.7	1.29e-4	1.83e-4
DIF322	0.0	-0.5	1.26e-4	1.79e-4
DIF323	0.7	-0.7	1.29e-4	1.83e-4
DIF332	0.7	0.0	1.16e-4	1.65e-4
DIF333	4.0	0.0	1.14e-4	1.61e-4

[4] Richard Soref and Brian Bennett, "Electrooptical Effects in Silicon", IEEE Journal of Quantum Electronics, Vol QE-23, January, 1987, p. 127

7.6 Results and Discussion

The phase change from a waveguide can be calculated from

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta nL \quad 7.14$$

However, if the index change in the waveguide is not uniform then the index changes within the waveguide must be summed. [5] Mathematically this can be expressed as

$$\Delta\phi = \sum_{l=1}^{m} \frac{2\pi}{\lambda}\Delta n_l \Delta L \quad 7.15$$

where

$\Delta\phi$ is the phase change,

λ is the wavelength,

Δn_l is the index change from x_l to x_{l+1},

and ΔL is the distance from x_l to x_{l+1} ($\Delta L = x_l - x_{l+1}$).

The predicted phase change was calculated from x=175 uM, the start of the gate region, to x=825 uM, the end of the gate region. At the center of the waveguide, y = 1.25 uM, the phase change vs. gate distance was graphed.

Graphs DIF321_LEN_PHASE is typical of a graph of the phase change vs. length for the DIF310 structure under different biasing conditions. A 180 degree phase shift is 3.1415 radians and the total phase change for the 650 uM waveguiding between the gate and substrate is significantly less. Although the phase changes are not as large as expected, it does give an indication of what biasing schemes are worth future consideration. The phase change is larger in the gate region but that region would not confine light.

Table 15: Summary of Phase Change Calculations in the Center of the Waveguide from x = 175 uM to x = 825 uM

title	anode (v3)	cathode (v4)	$\Delta\phi$(1.3)	$\Delta\phi$(1.55)	radians
DIF321	0.0	-0.7	0.0210	0.0248	
DIF322	0.0	-0.5	0.0196	0.0233	
DIF323	0.7	-0.7	0.0200	0.0237	
DIF332	0.7	0.0	0.0194	0.0231	
DIF332	4.0	0.0	0.0184	0.0218	

The five biasing schemes are representative of the types of conditions used to inject electrons and holes into a conduction channel. From these PISCES IIB simulations and optical calculations, one can draw the following conclusions:

- The peak current density for a longitudinal design will be less than a transverse one.

[5] P. C. Kendall, M. J. Adams, S. Ritchie, M. J. Robertson, "Theory for Calculating Approximate Values for the Propagation Constants of an Optical Rib Waveguide by Weighting the Refractive Indices", IEE Proceedings, Vol. 134, September, 1987, p. 701.

- The biggest obstacle is to source current from the anode without affecting the gate contact. Since the gate-cathode is a p-i-n junction, there is a limit of -0.5 volts that can be applied to the n+ cathode.
- The basic problem with a longitudinal design is that the current density in the conduction channel is not great enough to inject sufficient free carriers into the plasma to alter the index of refraction. Even though increasing the anode voltage will increase the current from the anode to the cathode, the applied anode voltage will cause the conduction channel to pinch off.
- While the PISCES IIB simulations predict changes in the electron and hole concentrations in the waveguiding region of the DIFET, they probably are not large enough for phase modulation applications.

CHAPTER 8

CONCLUSION

The motivation for this work was to analyze two dimensional rectangular silicon structures as electro-optic waveguides but the analysis method, can also be applied to any material and geometry. Although this work is concerned with only silicon, this approach can be adapted to simulate different semiconductor and insulator materials such as gallium arsenide (GaAs) or lithium niobate ($LiBN0_3$). PISCES IIB can be used to characterize the electrical properties of a semiconductor device under various doping or biasing conditions. This method included a more complete description of bimolecular recombination, diffusion currents and nonplanar current flow than previously employed. Index of refraction changes were calculated based on the change in the electron and hole concentrations under different biasing schemes. Large index changes within a confined optical region would mean that optical phase modulation is possible for that particular structure. Since many of the refractive index changes are not uniform across the waveguiding region of the device, the effective index change, Δn_{eff}, was also calculated. Silicon does not exhibit a linear optic effect, and has small second order electro-optic effects. Therefore, changes in its index of refraction due to charge injection were studied. Since silicon is a well understood material for electronics circuit design, it also holds the most promise for practical electro-optic waveguides

Using PISCES IIB, a more complete two dimensional model of the semiconductor structure has been presented. This model includes diffusion currents as well as recombination effects and two dimensional current flow. By injecting electrons and holes into the semiconductor plasma, changes in the real component of the refractive index can be induced. The refractive index changes are used as a basis for the design and analysis of electro-optic phase modulators.

Although it was not considered in this work, a Kramer-Kronig relation could be used to calculate the imaginary part of the dielectric constant from the real component of the index of refraction. With the imaginary component, the losses in the waveguide vs. distance could then be predicted.

Two different types of models were analyzed; double injection field effect transistors (DIFETs) and metal-oxide-semiconductor field effect transistors (MOSFETs). Although the DIFET electro-optic waveguide exhibits phase change potential, it is not as significant as the predicted changes in a one dimensional analysis. The problem with the DIFET is an inability to inject large enough electron and hole densities into the waveguide while maintaining optical confinement in the conduction channel. A large anode voltage, needed to inject holes, will pinch off the conduction channel and reduce the size of the optical waveguiding region. An anode voltage, needed to inject electrons, is limited because, below -0.5 volts, the gate region also becomes a source of holes. However, the holes are not injected into the waveguiding region but between the cathode and gate regions.

Both single and double injection MOSFET geometries were analyzed. Since the MOSFETs were transverse designs, the drain and source contacts were separated only by the width of the gate region. Consequently, the current density is the limiting factor in a transverse design. For insulated gate type structures, effective index changes on the order of 10^{-3} are possible but only in the depletion region beneath the insulated gate. Since the depletion region is dependent on concentration, larger refractive index changes are possible only in submicron waveguides.

8.1 Summary of Results

This work has studied four different geometries, MOS diode, single and double injection MOSFETs and longitudinal double injection FETs, for possible applications as silicon electro-optic phase modulators. The largest index changes are predicted in the simple MOS diode but it is also the structure with the smallest waveguiding region (only 0.25 uM thick). The peak index changes in the DIFET are positive because Δn_{max} is in the gate region not the waveguiding region.

Table 16: Summary of Maximum Refractive Index Changes for Geometries Studied

device	refer to section	injection	ΔV_g	$\Delta n(1.3)$	$\Delta n(1.55)$
MOS diode	2.5	none	-25.0	-4.68e-3	-6.64e-3
MOSFET	4.71	single	-4.4	-3.14e-4	-4.45e-4
MOSFET	5.52	double	-25.0	-1.43e-3	-2.03e-3
DIFET	7.5	double	-17.5	+1.29e-4	+1.83e-4

Comparing the MOSFET structures, larger effective refractive index changes are possible for double injection structures than for single injection. If the waveguiding area is thin enough (submicron) then 10^{-3} index changes can be realized.

Table 17: Summary of Maximum effective Refractive Index Changes for Transverse MOSFET Devices

device	injection	Area μM^2	ΔV_g	V_{ds}	$\Delta n_{eff}(1.3)$	$\Delta n_{eff}(1.55)$
MOSFET	single	1.2	-4.4	1.0	-1.68e-4	2.37e-4
MOSFET	double	24	-25.0	0.85	-1.96e-5	-2.78e-5
MOSFET	double	1.2	-25.0	0.85	-6.76e-4	-9.77e-4

8.2 Future Studies

Longitudinal structures will probably need to be insulated gate structures. The reason is that an oxide layer between the gate contact an the waveguide will provide a strong confinement layer. If the substrate region is also oxide, then the waveguide will be symmetrical because it will be bounded, top and bottom, by media with the same refractive index.

Waveguides with an insulated gate will also have to be submicron thick devices. The predicted waveguide thickness can be calculated by solving the equation for the maximum depletion width for the maximum injected electron or hole concentration.

Transverse waveguide structures are limited by the peak current density in the device. If the drain and source contacts are separated far enough apart then short channel effects can be ignored. However, for single mode waveguides, the maximum calculated separation distance is 12.0 uM.

In conclusion, this work certainly has not exhausted all the possibilities for electro-optic silicon waveguide design. Our study has developed the theory and technique for analyzing novel waveguide structures in both transverse and longitudinal configurations. Previously, only a one dimensional analysis of optical waveguiding structures was available and these analysis techniques were used to examine waveguide designs proposed in earlier work.

CHAPTER 9

SELECTED BIBLIOGRAPHY

Adams, M. J. An Introduction to Optical Waveguides, New York: Wiley & Sons, 1981.

Adams, M. J., S. Ritchie and M. J. Robertson. "Optimum Overlap of Electric and Optical Fields in Semiconductor Waveguide Devices." Applied Physics Letters, Vol. 48, 31 March 1986, 820-2.

Alferness, R. C. "Optical Guided-Wave Devices." Science, 14 November 1986, 825-9.

Blatt, Fred. Physics of Electronic Conduction in Solids, New York: McGraw-Hill, 1968.

Cap, F. F. "New Analytical 3D Method to Calculate Electromagnetic Waves in Glass Fibers of Arbitrary Cross Section Curvature." Integrated Optical Circuit Engineering III, SPIE Proceedings, vol. 651, Australia, 1986, 133-5.

Friedman, Lionel. "Proposal to Air Force Office of Scientific Research." Research Initiation Program, December, 1987.

Friedman, Lionel, Richard Soref and Joseph Lorenzo. "Silicon Double-Injection Electro-optic Modulators with Junction Gate Control." Journal of Applied Physics, 15 March 1988, 1831-9.

Hack, M., M. Shur and W. Czubatyj. "Double-Injection Field-Effect Transistor: A New Type of Solid-State Device." Applied Physics Letters, vol. 48, 19 May 1986, 1386-8.

Hartnagel, Hans. Semiconductor Plasma Instabilities, New York: American Elsevier Publishing, 1969.

Hecht, Eugene. Optics, Reading, MA: Addison-Wesley, 1987.

Hickernell, Fred. "Optical Waveguides on Silicon." Solid State Technology, November, 1988, 83-8.

Hockney, Roger, and James Eastwood. Computer Simulations Using Particles, New York: McGraw-Hill, 1981.

Holt, E. H. and R. E. Haskell. Foundations of Plasma Dynamics, New York: MacMillan, 1965.

Hutcheson, Lynn. Integrated Optical Cirucits and Components, New York: Marcel Dekker, 1987.

Integrated Optical Circuit Engineering VI, SPIE Proceedings, Vol. 993, Boston, MA, September, 1988.

Kendal, P. C., M. J. Adams, S. Ritchie and M. J. Robertson. "Theory for Calculating Approximate Values for Propagation Constants of an Optical Rib Waveguide by Weighting the Refractive Indices." IEE Proceedings, Vol. 134, September, 1987, 699-702.

Kittel, Charles. Introduction to Solid State Physics, New York: Wiley & Sons, 1986.

Lampert, Murray and Peter Mark. Current Injection in Solids, New York: Academic Press, 1970.

Levy, Ronald. "Integrated Photonic Devices on Silicon." Solid State Technology, November, 1988, 81.

Marcatilli, E. A. "Dielectric Rectangular Waveguide and Coupler for Integrated Optics." Bell System Technical Journal, vol. 48, September, 1969, 2072-89.

Marcuse, Dietrich. Theory of Dielectric Optical Waveguides, New York: Academic Press, 1974.

McCaughan, Leon. "Advanced Guided-Wave Integrated Optic Devices." Short course at the SPIE Conference, Boston, MA, September 6, 1988.

Moss, T. S. Optical Properties of Semiconductors, London: Buttersworth Scientific Publications, 1959.

Musikant, Solomon. Optical Properties, New York: Marcel Dekker, 1985.

Pankove, J. Optical Processes in Semiconductors, Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Pinto, Mark, Connor Rafferty and Robert Dutton. "PISCES-II User's Manual." Stanford University: Stanford Electronics Laboratories, 1984.

Quimby, Richard. Introduction to Optoelectronics lecture notes, Worcester Polytechnic Institute, January, 1988.

Seraphin, B. O. and N. Botka. "Franz-Keldysh Effect of the Refractive Index in Semiconductors." Physics Review, Vol. 139, 19 July 1965, 562-5.

Sluss, James. "An Introduction to Integrated Optics for Computing." Computer, December, 1987, 9-23.

Soref, Richard. "Notes on Silicon Triodes." Private memo, January 19,1988.

Soref, Richard. Private memo to Prof. Lionel Friedman, 11 May 1988.

Soref, Richard and Brian Bennett. "Electrooptical Effects in Silicon." IEEE Journal of Quantum Electronics, January, 1987, 123-8.

Soref, Richard and Joseph Lorenzo. "1.3 uM Electro-optic Silicon Switch." Applied Physics Letters, Vol. 51, 6 July, 1987, 6-8.

Soref, Richard and Joseph Lorenzo. "All-Silicon Active and Passive Guided-Wave Components for 1.3 uM and 1.6 uM Wavelengths." IEEE Journal of Quantum Electronics, June, 1986, 873-9.

Streetman, Ben. Solid State Electronic Devices, Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

Suematsu, Yasuharu, and Ken-Ichi Iga. Introduction to Optical Fiber Communication, New York: Wiley & Sons, 1982.

Sze, S. M. Semiconductor Devices Physics and Technology, New York: Wiley & Sons, 1985.

Teit, M. D. and J. A. Fleck, "Light Propagation in Graded-Index Optical Fibers." Applied Optics, Vol. 17, 15 December 1978, 3990-8.

Whalen, M. S. and J. Stone. "Index of Refraction of n-type InP at 0.6633-um and 1.15 um Wavelengths as a Function of Carrier Concentration." Journal of Applied Physics, June, 1982, 4340-3.

Wilson, J. and J. Hawkes. Optoelectronics: An Introduction, Englewood Cliffs, New Jersey: Prentice-Hall, 1983.

Yariv, Amnon and Pochi Yeh. Optical Waves in Crystals, New York: Wiley & Sons, 1984.

Young, T. P., G. A. Armstrong and W. P. Smith. "Semiconductor Finite Elements for Integrated Optics." GEC Research Ltd., November, 1987, 1-6.

Appendices can be obtained from
Universal Energy Systems, Inc.

1986 USAF-UES MINI GRANT PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

UNIVERSAL ENERGY SYSTEMS, INC.

MINI-GRANT FINAL REPORT

MEASUREMENTS OF A SLOT ANTENNA FED BY COPLANAR WAVEGUIDE AND SOLUTION OF AN INFINITE PHASED ARRAY OF SLOTS FED BY COPLANAR WAVEGUIDE OVER A DIELECTRIC HALF-SPACE

Prepared by:	Dr. Donald F. Hanson
Academic Rank:	Associate Professor
Department and University:	Department of Electrical Engineering University of Mississippi University, MS 38677
USAF Research Contact:	Dr. Robert J. Mailloux Rome Air Development Center EEA Hanscom AFB, MA 01731
Date:	June 1, 1988
Time period of grant:	Jan. 15 - Dec. 15, 1987

I. INTRODUCTION

A. Work Completed

The Air Force has a vital interest in the rapidly developing Monolithic Microwave Integrated Circuit (MMIC) technology. Interest in developing MMIC integrated circuit antennas for phased arrays is high. One type of antenna element that shows great promise for this application is the slot antenna fed by coplanar waveguide.

This writer's USAF-UES Summer Faculty Research Program (SFRP) project was to study the "Fields of a Slot Antenna on a Half-Space Fed by Coplanar Waveguide Using the Method of Moments". This writer's final report [1] from 1986 provides a numerical (Moment Method) solution to the single slot fed by coplanar waveguide. This has been reported [2] in the literature. This writer's final report [3] from 1985 provides (a) a literature review on the topic of coplanar waveguide; (b) applicable source configurations for modeling slot receivers/radiators in coplanar waveguide; and (c) mathematical methods for formulating planar phased arrays fed by coplanar waveguide. The mini-grant proposal [4] outlined three objectives for 1987. First, design data for a single slot antenna fed by coplanar waveguide were to be obtained. Second, design data for an infinite phased array of slots were to be obtained. Finally, another objective was to do measurements of a single slot antenna fed by coplanar waveguide. The last two objectives, being new, were the first to be worked on. In fact, the third objective on measurements became a major commitment in time and effort. At this time, the computer program for the second objective (arrays) has been finished, but design data have not yet been obtained.

Preliminary measurements for several cases are complete and more are being done. Therefore, this final report contains major sections on (1) measurements of slot antennas fed by coplanar waveguide and (2) integral equations for an infinite phased array of slots fed by coplanar waveguide. This work will continue. A Cyber 860/180 computer and a Cyber 205 supercomputer have recently arrived on campus at the University of Mississippi. The programs which I wrote during 1986 at Hanscom AFB will run without modification on the Cyber 860/180. The Cyber 205 supercomputer will be very useful to solve the infinite phased array case.

B. Coplanar Waveguide

Coplanar waveguide uses the transmission line structure shown in Figure 1. The two outer half-planes are grounded and the signal is fed in on the center line. The thin metal structure is held in place by dielectric, which is a half-space ($y<0$) of dielectric constant $\epsilon_r \epsilon_0$ for the numerical calculations. Two slots are present for $a < |x| < b$. The x-extent of the metal is $0 \leq |x| < a$ and $|x| > b$.

Coplanar waveguide (CPW) is used in many circuit applications because components can be mounted in either series or shunt. It is also especially convenient for three terminal active devices like FETs which require both shunt and series connections at the same location. The ground plane metalization then surrounds the device which makes low inductance connections to ground possible. This gives increased gain in an amplifier configuration. CPW also provides reliable microwave frequency interconnections on semiconductor surfaces for monolithic semiconductor integrated circuits. All microwave frequency interconnections can be made on a single coplanar surface. Recent work

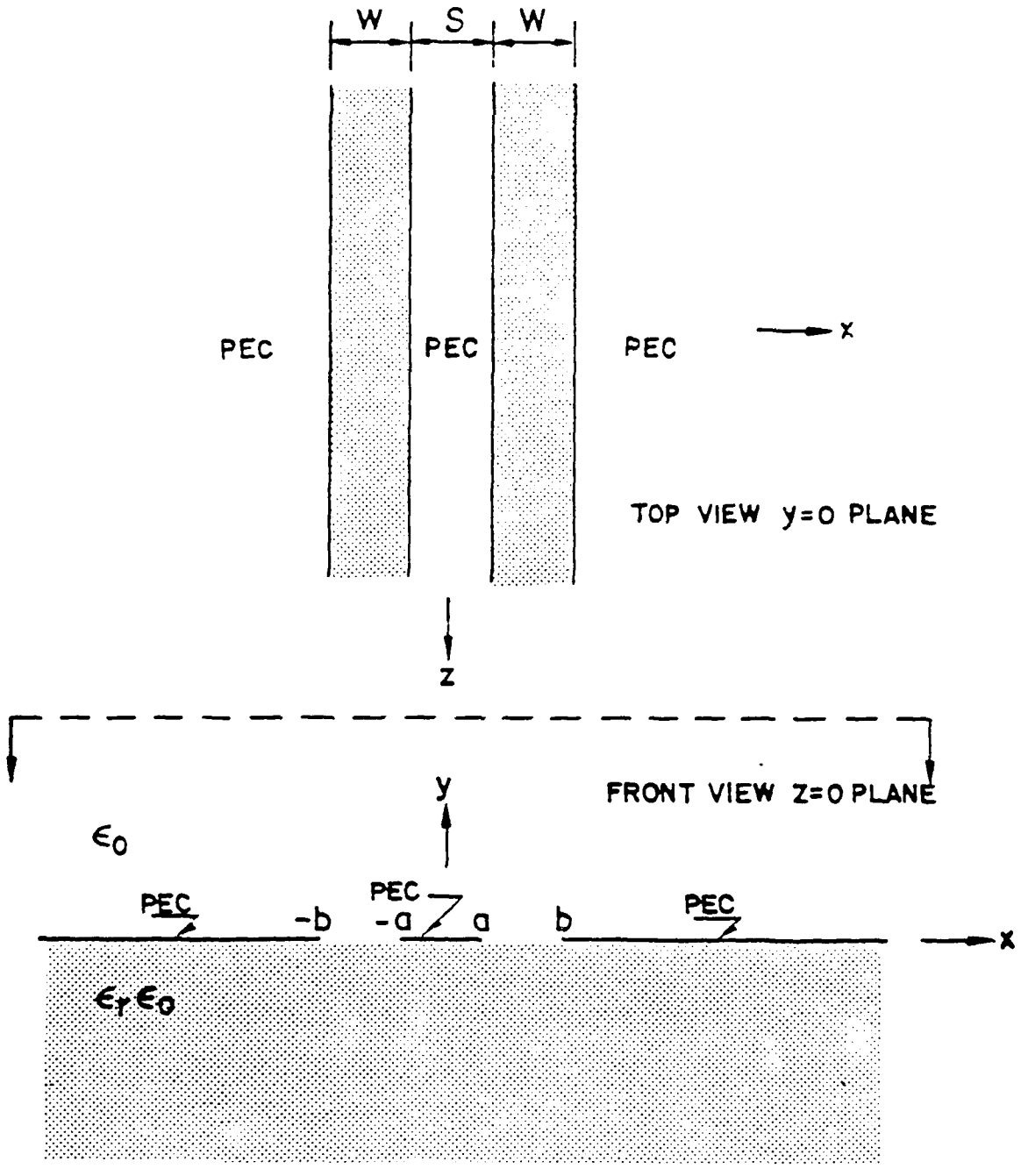


Figure 1. Coplanar Waveguide Geometry.

in slow-wave CPW structures has provided the ability to minimize the chip surface area required for transmission line applications in exchange for increased attenuation down the line. Coplanar inductors and power FETs can be made using air bridge technology. Reliable methods for routing DC biasing lines over the CPW to the active device sites must still be developed, however.

C.P. Wen [5] apparently first described CPW and calculated its characteristic impedance. Using conformal mapping techniques, he found it to be

$$Z_0 = \frac{\sqrt{\mu_0/\epsilon_0}}{4} \frac{1}{\sqrt{\frac{1+\epsilon_r}{2}}} \frac{K(k')}{K(k)} \tag{1.1}$$

$$k = \frac{a}{b}, \quad k' = \sqrt{1 - \left(\frac{a}{b}\right)^2} .$$

where K is the complete elliptic integral of the first kind. An excellent review of the pre-1979 literature on CPW is found in [6]. The most recent literature on CPW applications, particularly relating to active devices, is described in [3]. A recent text [7], Applications of GaAs MESFETs, covers the use of coplanar waveguide on GaAs substrates. Pucel [8] describes the use of CPW in monolithic microwave circuits and compares it with microstrip, coplanar strips, and slot line. He concludes that of the four, microstrip and CPW are best for GaAs monolithic circuits, but that microstrip is preferred

C. Antennas

Antennas proposed for incorporation into planar phased arrays are the microstrip dipole or patch antennas, and the slot dipole or ring antennas. A recent (1983) book chapter "Integrated-Circuit Antennas" [9] covers the applicable concepts. Another book, Microstrip Antennas [10], surveys microstrip and slot antennas and is a handy reference. A special issue of IEEE Trans. Antennas Prop. [11] on microstrip antennas covers many aspects of the area.

The substrate that is currently most often discussed for monolithic antennas is GaAs which has a relative dielectric constant of approximately 13. This causes problems when antennas are mounted on its surface. Large amounts of power can be trapped inside the substrate instead of being radiated into space. Reciprocity is often used to calculate radiation patterns of microstrip and slot antennas.

During the summer of 1986, a computer program was written for a Cyber computer to solve for the fields of a single slot antenna fed by coplanar waveguide over a dielectric half-space. The integral equation formulation for this case is given in [1]. The integral equation is solved using the moment method. Results are given in [1] and [2]. Measurements of such a single slot antenna fed by coplanar waveguide have been made and are described in this report. The integral equation formulation for an infinite phased array of slot antennas fed by coplanar waveguide is also given in this report.

II. MEASUREMENTS OF SLOT ANTENNAS FED BY COPLANAR WAVEGUIDE

The measurements were of two types, input impedance and radiation patterns. Figure 2 shows a slot antenna fed by a section of Coplanar Waveguide (CPW). The input impedance of the slot needs to be measured at the intersection of the slot and the CPW. Therefore, the impedance measurement at the connector must be corrected to obtain the proper impedance at the slot antenna. To do this, a two-port section of CPW was measured to obtain the de-embedding data.

Printed circuit substrate materials obtained from RADC were of two types. Oak 602 Teflon and 3M EPSILAM-10, hereafter referred to as Type A and Type B substrates, respectively. Type A material was 1/16" thick and has a relative dielectric constant ϵ_r of 2.54. Type B material was 1/20" thick with $\epsilon_{rzz} = 10.2$, $\epsilon_{rxx} = \epsilon_{ryy} = 15.0$. Antennas fed by CPW were made from both substrates. All connectors obtained from RADC were of OSM 215-3CCSF SMA type. These were the materials used throughout. Both the grounded, or copper-backed, case and the ungrounded, or plain-bottomed, case are examined. Figure 3 shows the geometry for the grounded case.

This section on measurements is divided into several parts: First, there is the design of the CPW. Then, there is the design of the slot antennas. Third, there is the theory of connector and feedline characterization. Fourth, there is the measurement of the CPW two-port sections. Fifth, there is the de-embedding of the input impedance and determination of the resonant frequency. Finally, the far-field patterns are measured.

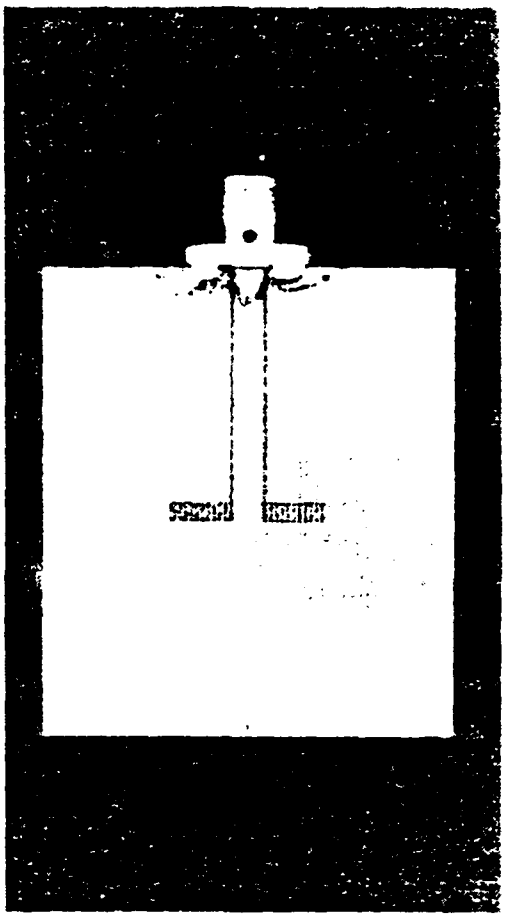


Figure 2. Slot Antenna Fed by CPW with SMA Connector.

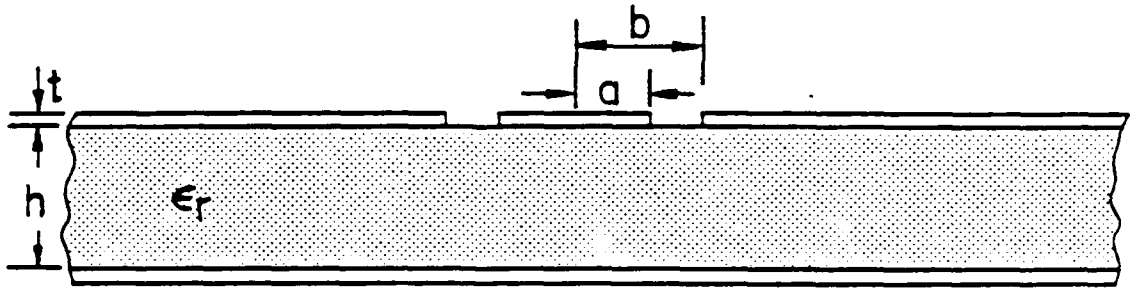


Figure 3. Grounded or Copper-Backed Coplanar Waveguide.

A. Design of CPW

First, the CPW feedline to the slot antenna needs to be designed. Since the measurement equipment is based on 50Ω, the design value for the CPW characteristic impedance was chosen to be $R_0 = 50\Omega$. For each substrate material, both the grounded and the ungrounded cases were designed. These are:

(1) Type A substrate, grounded
(2) Type A substrate, ungrounded
(3) Type B substrate, grounded
(4) Type B substrate, ungrounded

The design used a computer program from [12]. For the ungrounded case, the effective ϵ_r [13] that was used was

$$\epsilon_{reff} = 1 + q(\epsilon_r - 1) \tag{2.1}$$

where

$$q = \frac{1}{2} \frac{K(k_1)/K(k_1')}{K(k)/K(k')} \tag{2.2a}$$

$$k = \frac{a}{b} \tag{2.2b}$$

and

$$k_1 = \sinh(\pi a/2h)/\sinh(\pi b/2h) \tag{2.2c}$$

Figure 3 shows the dimensions for a, b, and h. K(k) is the complete elliptic integral of the first kind and $k_i' = (1 - k_i^2)^{1/2}$. The characteristic impedance is given by

$$Z_u = \frac{30\pi}{\sqrt{\epsilon_{reff}}\, K(k)/K(k')} \tag{2.3}$$

These formulas were modified to account for the cladding thickness t using formulas from [14].

For the grounded case, a formula from [15] was used. The grounded case is treated as a combination of a microstrip and an ungrounded CPW.

Therefore, the formula uses both the previous formula for Z_u and a formula Z_m for the characteristic impedance of microstrip. The grounded case formula [15] is

$$Z = \left[\frac{5q}{1+5q} \frac{1}{Z_m} + \frac{1}{1+q} \frac{1}{Z_n} \right]^{-1} \qquad (2.4a)$$

where $$q = \frac{2a}{h} \left(\frac{b}{a} - 1 \right) \left\{ 3.6 - 2.0 \exp \left[-(\epsilon_r + 1)/4 \right] \right\} \qquad (2.4b)$$

A formula from [14, pp. 62-63] was used to compute Z_m, including the effect of the cladding thickness t.

These formulas for Z_u and Z_g were used to design the coplanar waveguide feedlines. A reasonable value of strip width S = 2a was chosen and then slot width W = b-a was varied until a value close to 50Ω was obtained. For the cases 1 through 4 given at the beginning of this section, the design values for W and S and the computed values of R_0 are given in Table 1 below.

Table 1. CPW Design.

Case >	1	2	3	4
S	3.0 mm	3.0 mm	1.0 mm	1.8 mm
W	0.5 mm	0.25 mm	1.9 mm	1.0 mm
R_0	Z_g = 50.5Ω	Z_u = 51.2Ω	Z_g = 49.5Ω	Z_u = 50.5Ω
Type	A	A	B	B

B. Design of Slot Antennas

The design of the slot antenna was done by determining an effective dielectric constant and then finding the length corresponding to the wavelength divided by two in the medium. Two cases were designed. First, the case of a slot fed by CPW without conductor backing, and second, the case with conductor backing.

For the case without conductor backing, it was decided to use an ϵ_e expression similar to that used in the design of the CPW even though an expression for slot lines may be more correct physically. An appropriate slot line expression was not found, however. Therefore, ϵ_e was computed from the following equation [6, p. 275]

$$\epsilon_{reff} = \frac{\epsilon_r + 1}{2} \left[\tanh\,(1.785\,\log(h/W) + 1.75) + \right.$$

$$\left. + \frac{kW}{h}\,(0.0\ldots - 0.7k + 0.01\,(1.0 - 0.1\,\epsilon_r)(0.25 + k)) \right] \qquad (2.5)$$

where $k = \frac{a}{b} = S/(S + 2W)$ and h is the substrate thickness.

For the case of a slot antenna fed by CPW with conductor backing, as shown in Figure 3, ϵ_{reff} was computed as the average between the ϵ_{reff} of the CPW feedline and the ϵ_{reff} of the slot antenna when viewed as a section of CPW. This was done because formulas for ϵ_{reff} of the slot antenna slot line could not be found. Therefore, the two ϵ_{reff} formulas were computed by a computer program "EPSLON" using the following formulas [16]:

$$\epsilon_{reff} = 1 + q(\epsilon_r - 1) \qquad (2.6a)$$

where

$$q = \frac{K(k_1)/K(k_1')}{K(k_1)/K(k_1') + K(k)/K(k')} \qquad (2.6b)$$

$$k = \frac{a}{b} \qquad (2.6c)$$

and

$$k_1 = \tanh(\pi a/2h)/\tanh(\pi b/2h). \qquad (2.6d)$$

The values of a, b, and h are shown in Figure 3.

The resonant length of the slot antennas was chosen to be

$$L/\lambda_0 = 0.485 \frac{1}{\sqrt{\epsilon_{reff}}} \tag{2.7}$$

for all cases. This choice is similar to

$$\frac{L}{\lambda_0} = \frac{0.48}{\sqrt{\epsilon_e}} \frac{1}{1 + \frac{W}{L}} \tag{2.8}$$

given in [17].

For a nominal resonant frequency of 6 GHz, the computed values of ϵ_{reff} and L are shown in Table 2.

Table 2. Antenna Lengths f = 6 GHz

Antenna >	1	2	3	4
ϵ_{reff}	1.88	1.50	6.59	4.85
L(mm)	17.70	19.80	9.45	11.00

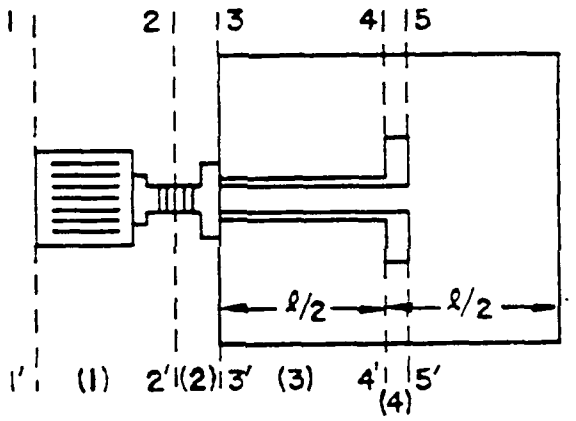
$$L = 50(.485)/\sqrt{\epsilon_{reff}} \text{ mm}$$

Since experimental resonant frequencies of this length were found to be 6% too high and since it can be shown that dL/L = -df/f, antennas 1, 3, and 4 are being remade with a length L' = 1.06L.

The printed circuit board layouts were prepared using the computer-aided-drafting program AutoCAD. On examining the finished boards on a laboratory microscope, they were found to be quite close to the desired dimensions. Figure 2 shows such a finished antenna.

C. Connector and Feedline Characterization

The input impedance to the slot antenna fed by CPW must be measured accurately at the input which is the line 4-4', as shown in Figure 4. Therefore, it is necessary to characterize the CPW feedline (3), the SMA/PCB connector (2), and the APC7/SMA adapter (1), as shown in



(1) APC7/SMA adaptor

(2) SMA/PCB connector

(3) CPW feedline

(4) slot antenna

Figure 4. Slot Antenna Fed by Coplanar Waveguide. with Connector and Adapter

Figure 4. The APC7/SMA adapter is necessary to connect the board to the reflection/transmission test set, a component of the network analyzer system. If the devices (1), (2), and (3) are characterized, then input impedance measurements made at 1-1' can be used to de-embed the input impedance at 4-4'.

To characterize the devices (1), (2), and (3), a method suggested by [18] was employed. A section of CPW was needed that was identical to the CPW line on the slot antenna board, except twice the length $\ell/2$ where $\ell/2$ is shown in Figure 4. Identical connectors and adapters were needed on each end so that the board is symmetric from left to right. Such a CPW two-port is shown in Figure 5. The idea [18] is that the precision APC7/SMA connectors (devices (1) and (1)') and the length of CPW line (3) are modeled as simple transmission line sections. The discontinuities in sections (2) and (2)' are modeled by S parameters. The S matrices for both sections (2) and (2)' are assumed to be identical. The signal flow graph [19] for the configuration shown in Figure 5 is shown in Figure 6.

Using the CPW design data given in Table 1, four 2-inch sections of CPW were built, as shown in Figure 7. The layout for the boards was done using AutoCAD computer-aided-drafting software on an IBM-PC. Some small errors were encountered since the laser printer had only 75 line/inch resolution in graphics mode. The layouts were made 4 times larger than actual size, so line widths referenced to actual size were quantized in steps of 0.085 mm. For this reason and also because of the manufacturing tolerances in the etching equipment, the measured dimensions are not exactly the same as designed. The measured dimensions are shown in Table 3 below. Note that the two slots were of slightly different widths W on board 1 and board 4.

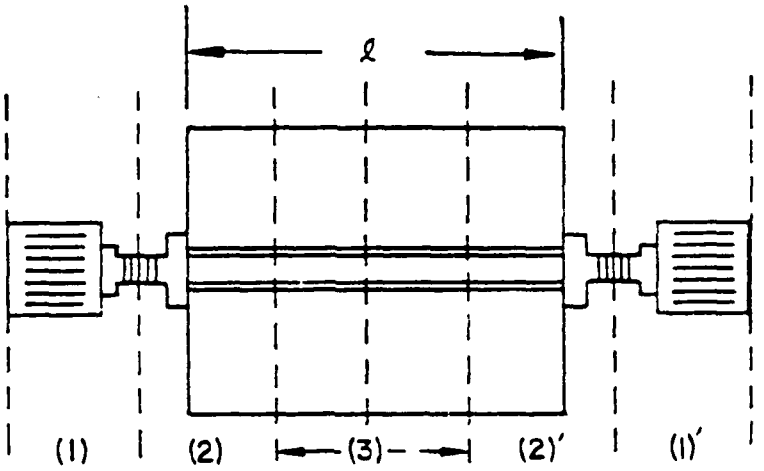


Figure 5. 50Ω Coplanar Waveguide and Coaxial Connectors.

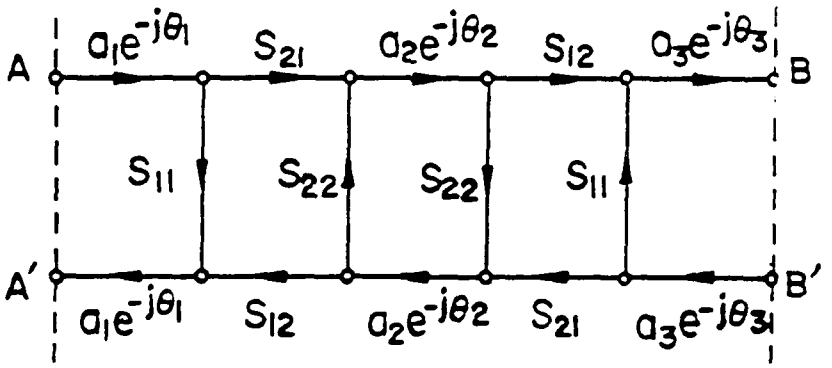


Figure 6. Signal Flow Graph Modeling the Configuration of Figure 5.

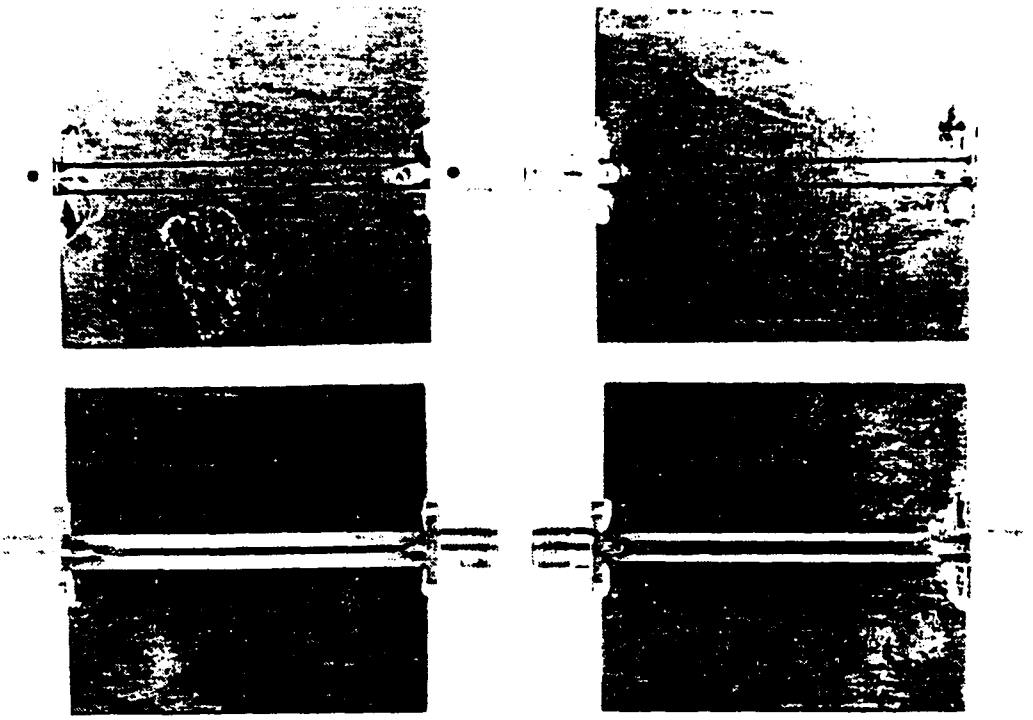


Figure 7. Two Inch Sections of CPW.
The Left Two are Copper-Backed.
The Right Two are Plain-Bottomed.

Table 3. Measured Dimensions.

Board	1	2	3	4
S	3.0 mm	2.9 mm	0.9 mm	1.9 mm
W	0.55/ 0.58 mm	0.38 mm	2.0 mm	1.1/ 1.0 mm

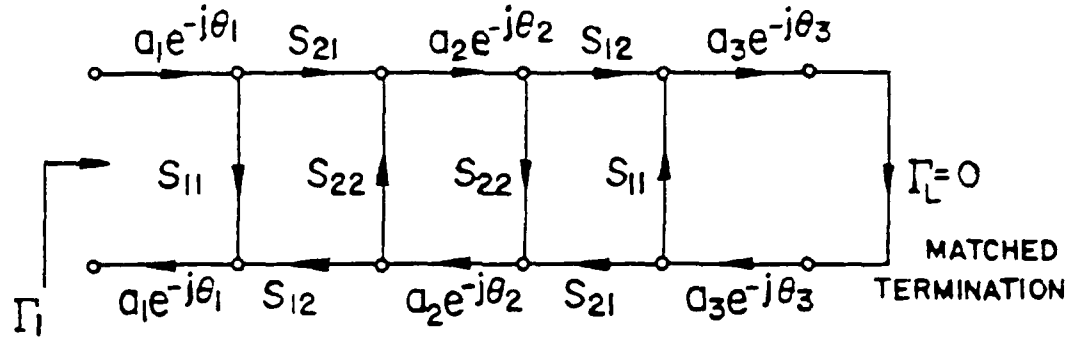
The four CPW sections were used to obtain accurate de-embedding data for the slot antenna fed by CPW input impedance. Referring to Figure 6, it will be assumed that $S_{21} = S_{12}$ and that a_1, a_2, and a_3 and θ_1, θ_2, and θ_3 are known. Therefore, the three remaining unknowns are S_{11}, S_{22}, and $S = S_{21} = S_{12}$. Three measurements of the CPW boards are necessary to determine these values. The three calibration measurements that were performed are reflection measurements with matched termination and short circuit termination and a transmission measurement.

First applying a matched termination to the CPW board, the signal flow graph is as shown in Figure 8(a). Using Mason's gain formula with approximation [19], the reflection coefficient for the matched termination becomes

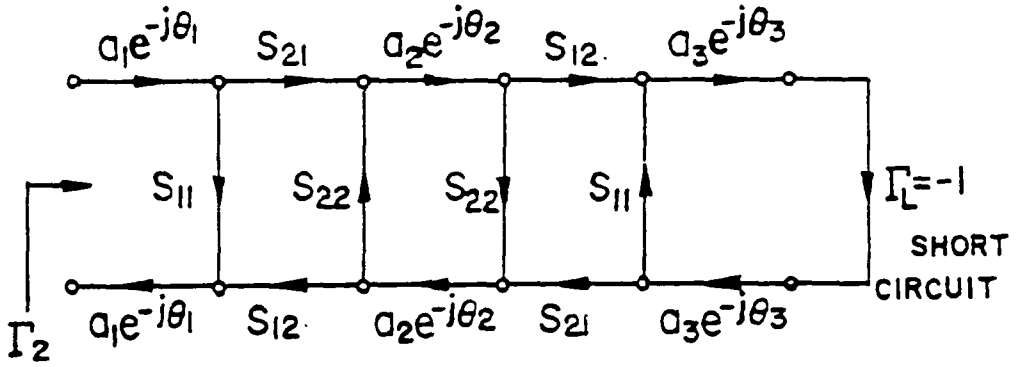
$$\Gamma_1 = a_1^2 S_{11} e^{-j2\theta_1} + a_1^2 a_2^2 S_{12}^2 S_{22} e^{-j2(\theta_1+\theta_2)} \qquad (2.9a)$$

This assumes $|S_{22}| << 1$. Second, a short circuit termination is applied to the CPW. The signal flow graph model becomes as shown in Figure 8(b). Using Mason's gain formula and assuming $|S_{22}| << 1$, the reflection coefficient becomes

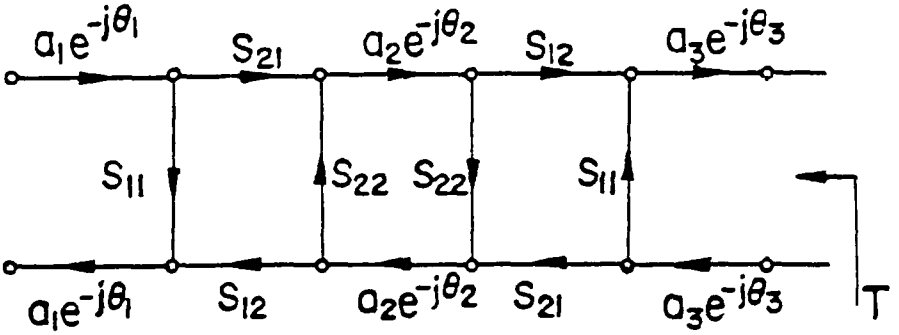
$$\Gamma_2 = a_1^2 S_{11} e^{-j2\theta_1} + a_1^2 a_2^2 S_{12}^2 S_{22} e^{-j2(\theta_1+\theta_2)} - a_1^2 a_2^2 a_3^2 S_{12}^4 e^{-j2(\theta_1+\theta_2+\theta_3)} (1 - a_3^2 S_{11} e^{-j2\theta_3}) \qquad (2.9b)$$



(a)



(b)



(c)

Figure 8. Configurations for Calibration Measurements.
(a) Reflection (Matched).
(b) Reflection (Short Circuit).
(c) Transmission.

Finally, a transmission measurement is used. The signal flow graph for the transmission model is shown in Figure 8(c). Again, applying Mason's gain formula and assuming $|S_{22}| \ll 1$, one obtains

$$T = a_1 \, a_2 \, a_3 \, S_{12}^2 \, e^{-j(\theta_1+\theta_2+\theta_3)} \tag{2.9c}$$

Once Γ_1, Γ_2, and T are known, the values of $S = S_{12} = S_{21}$, S_{11} and S_{22} can be found by simultaneous solution. The values are

$$S_{12}^2 = \frac{T}{a_1 \, a_2 \, a_3} \, e^{j(\theta_1+\theta_2+\theta_3)} \tag{2.10a}$$

$$S_{11} = \frac{T^2 + \Gamma_2 - \Gamma_1}{T^2 \, a_3^2 \, e^{-j2\theta_3}} \tag{2.10b}$$

$$S_{22} = \frac{\Gamma_1 - a_1^2 \, S_{11} \, e^{-j2\theta_1}}{a_1^2 \, a_2^2 \, S_{12}^2 \, e^{-j2(\theta_1+\theta_2)}} \tag{2.10c}$$

In this case of a slot antenna fed by CPW line, the section of line identified by (3) in Figure 5 was chosen to be of zero length. This choice required that the CPW line section be made of length ℓ in Figure 5 where $\ell/2$ is the length of the CPW feedline section shown in Figure 4. Thus, both the CPW feedline and the SMA/PCB connector may be characterized by the S-parameters. Thus, $a_2 = 1.0$ and $\theta_2 = 0.0$. The values of a_1, a_3, and θ_1, θ_3 were determined from four independent measurements of the APC7/SMA adapters. In the range of frequencies from 5 to 7 GHz, a_1 and a_3 were found to be almost 1.0. Values for θ_1 and θ_3 were obtained.

Using the assumption that $a_2 = 1.0$ and $\theta_2 = 0.0$, the signal flow graph for the slot antenna fed by CPW is as shown in Figure 9. Using Mason's gain formula, one obtains

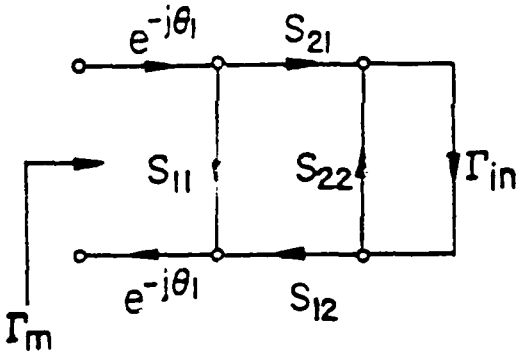


Figure 9. Signal Flow Graph for Slot Antenna Fed by CPW.

$$\Gamma_{in} = \frac{\Gamma_m - S_{11}\, e^{-j2\theta_1}}{S_{12}^2\, e^{-j2\theta_1} + S_{22}\,(\Gamma_m - S_{11}\, e^{-j2\theta_1})} \qquad (2.11)$$

This is the formula for de-embedding the input impedance to the slot antenna fed by CPW.

D. Measurement of CPW Boards for S Parameter Characterization

Prior to making measurements of the CPW sections, measurements were made to assess the performance of the automatic network analyzer system which is illustrated in Figure 10. To do this, a precision 3dB attenuator was chosen to be the device under test. Both reflection and transmission type measurements were performed. For a reflection measurement with a load of the 3dB attenuator with short circuit termination, it was found that in the test frequency range between 5.0 and 7.0 GHz, a reflection coefficient of about 6dB was obtained as expected. The magnitudes of S_{11} and S_{22} for the 3dB attenuator differed by less than 0.05 dB and the phase of S_{11} and S_{22} differed by less than 1°. For the transmission measurement, it was found that the transmission coefficient measured had a magnitude very close to 3dB.

Next, a measurement was performed on the APC7/SMA precision adapters (Sections (1) and (1)' of Figure 5). The adapters on hand were not identical and were of slightly different lengths. Their transmission coefficients can be expressed as $a_1\, e^{-j\theta_1}$ and $a_3\, e^{-j\theta_3}$, respectively. Over the range of frequencies from 5 to 7 GHz, empirical formulas were developed from the measured data. These formulas are

$$a_1 = a_3 = 1.0 \qquad (2.12a)$$

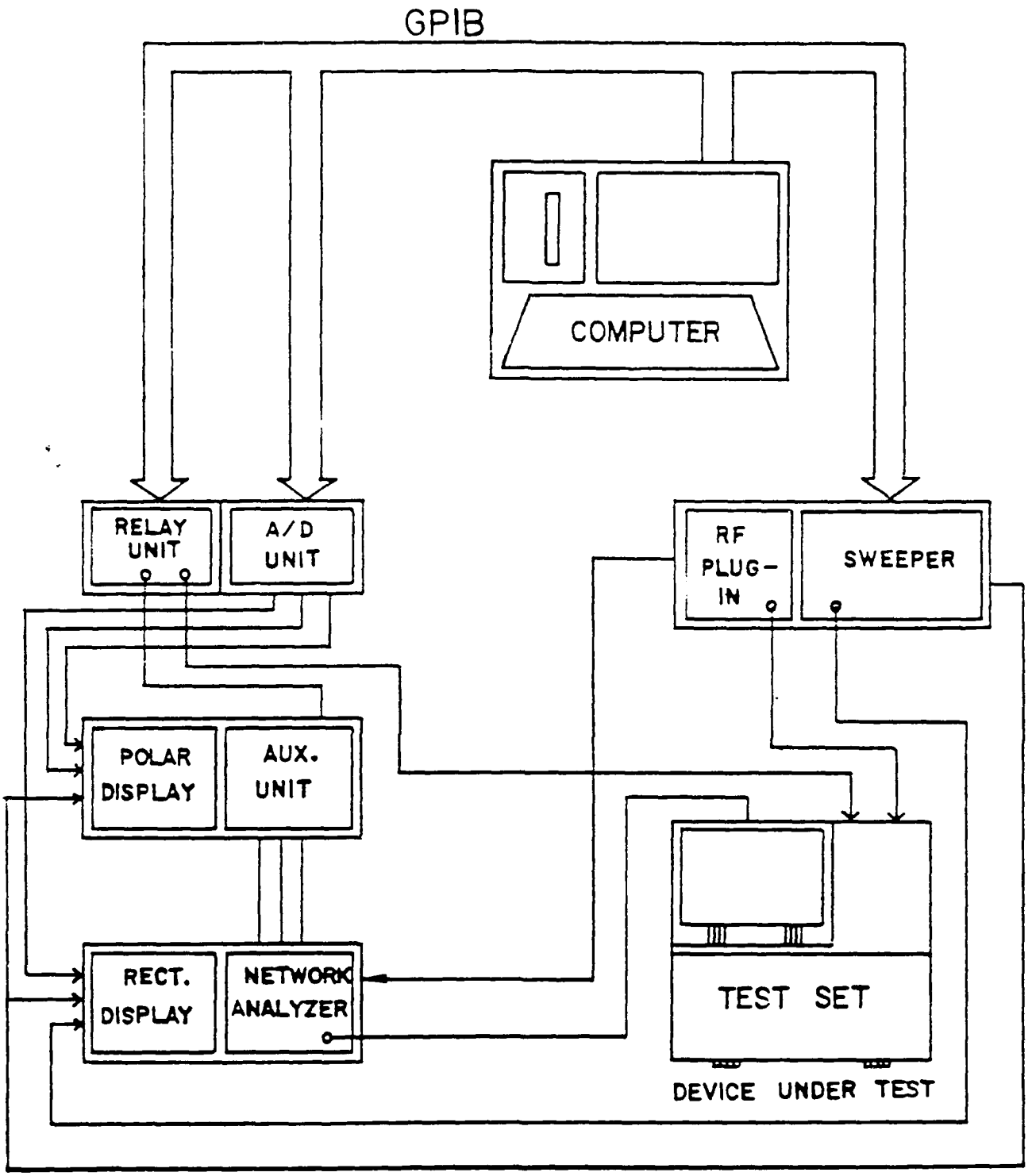


Figure 10. Automatic Network Analyzer System.

$$\theta_1 \text{ in degrees} = \begin{cases} 51.9\ (f - 6.4) + 301.9 & 5.0 \le f \le 6.4 \\ 37.8\ (f - 6.4) + 301.9 & 6.4 \le f \le 7.0 \end{cases} \quad (2.12b)$$

$$\theta_3 \text{ in degrees} = \begin{cases} 49.0\ (f - 6.4) + 286.1 & 5.0 \le f \le 6.4 \\ 34.7\ (f - 6.4) + 286.1 & 6.4 \le f \le 7.0 \end{cases} \quad (2.12c)$$

These formulas are used in the computation of the S parameters for the CPW feedline/connector combination, and are needed to de-embed the input impedance of the slot antenna.

As discussed previously, two reflection measurements, one with matched termination and the other with short circuit termination, and one transmission measurement need to be made of the CPW two-port situation shown in Figure 5 to determine the S parameters for the CPW line section and connector, as shown in Figures 4 and 9. Figure 11 shows a side-by-side photograph of the slot antenna board and the CPW two-port board for Case 1 of Table 1. Note that the CPW two-port board is twice as long as the section of CPW on the slot antenna board.

Initially, the two reflection measurements and one transmission measurement were performed on CPW two-port boards 2 and 4 described by Table 1. Later, this data was measured for board 1. This data was stored in a computer file for use in obtaining the de-embedded input impedance of the slot antenna. The data was used to compute the values for S_{11}, S_{22}, and S using Equations (2.10a-c).

E. Reflection Measurement of Slot Antenna

The experimental set-up for reflection coefficient measurement of a slot antenna board is the same as that shown in Figure 10. First, the network analyzer system is calibrated using three standard SMA

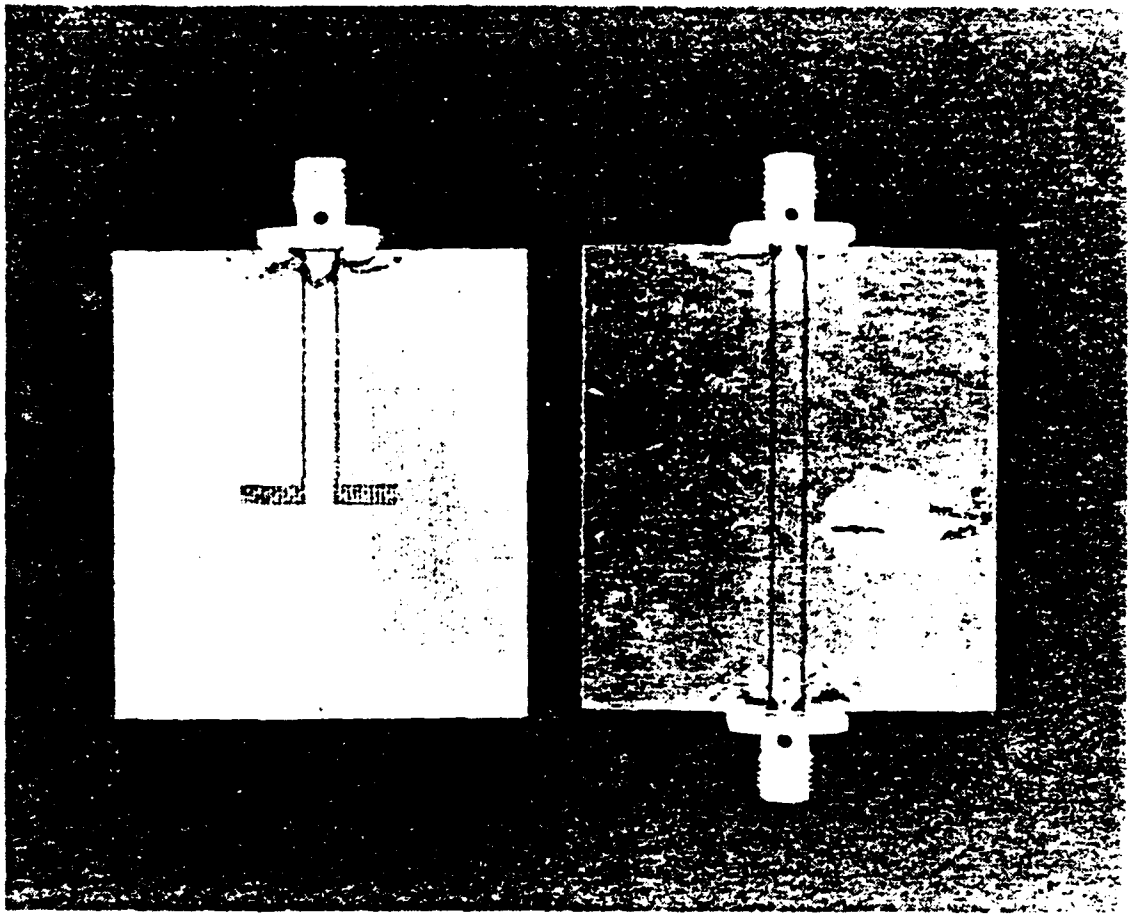


Figure 11. Slot Antenna Board with CPW Board.

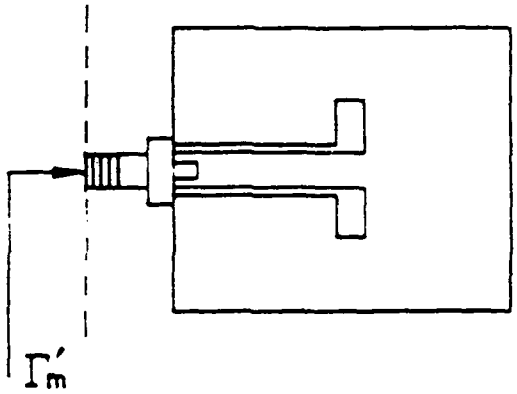
terminations - a matched 50Ω load, a short circuit, and an open circuit. Then, the reflection coefficient of the slot antenna board is measured, automatically error corrected, and the data are displayed on the computer display. The reference plane for the reflection coefficient data is defined as shown in Figure 12(a). Including the effects of the adapter, it is easy to make the reference plane, as shown in Figure 12(b). Equations (2.10) and (2.11) assume the reference plane is as shown in Figure 12(b).

The actual reflection coefficient Γ_L of the slot antenna is defined at the reference plane, as shown in Figure 13. A computer program "DMBED" reads the input data file CALIB and computes the S-parameters of the connector/CPW feedline section from Eq. (2.10). It also accepts the measured reflection coefficient data Γ_m and then computes $\Gamma_L = \Gamma_{in}$ defined, as shown in Figure 13, from Γ_m and [S] using Equation (2.11). The input impedance of the slot antenna is computed using

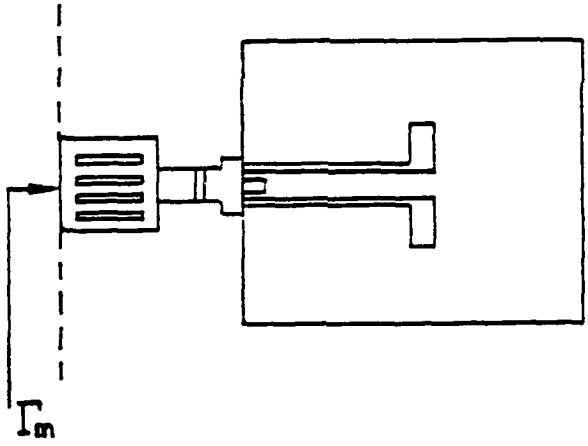
$$Z_{in} = \frac{1 + \Gamma_L}{1 - \Gamma_L} Z_0 \qquad (2.13)$$

where Z_0 is the characteristic impedance of the CPW feedline.

Preliminary measurements were made for boards 2 and 4 of Table 1. The measurements for board 4 were difficult to interpret and so initially work was focused on boards 1 and 2 which both were of Type A. Calibration measurements were made on board 1. These measurements were stored on the computer for use in the de-embedding scheme. The computer program outputs the de-embedded input admittance versus frequency data. Figures 14 and 15 show the de-embedded input admittance Y_{in} versus frequency for boards 1 and 2, respectively. The design resonant



(a)



(b)

Figure 12. Reference Planes for Reflection Coefficient Data.
(a) At Connector Input.
(b) At Adapter Input.

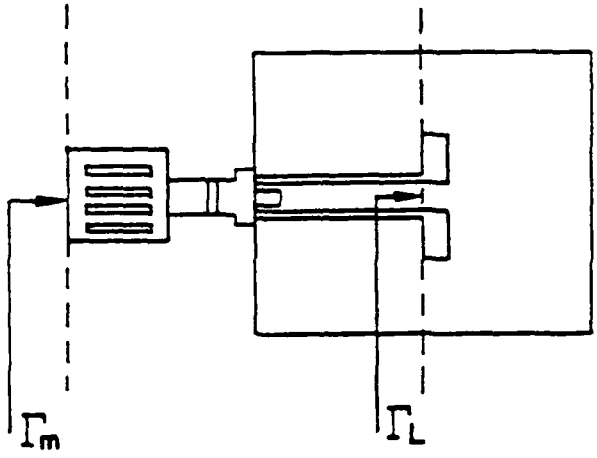
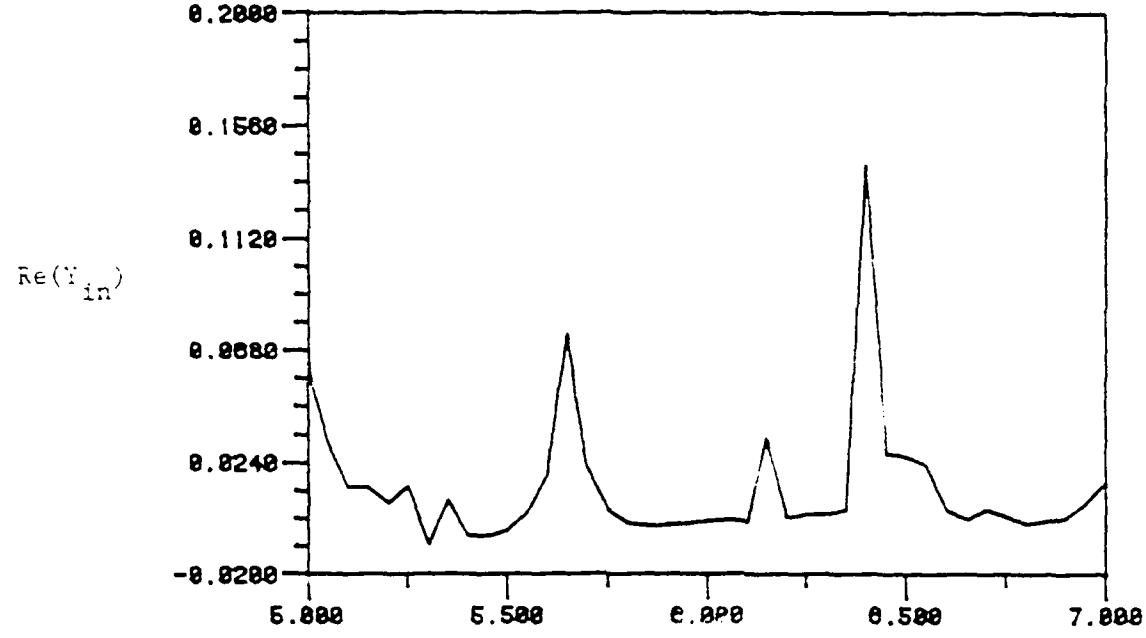


Figure 13. Reference Planes for Γ_L and Γ_m.



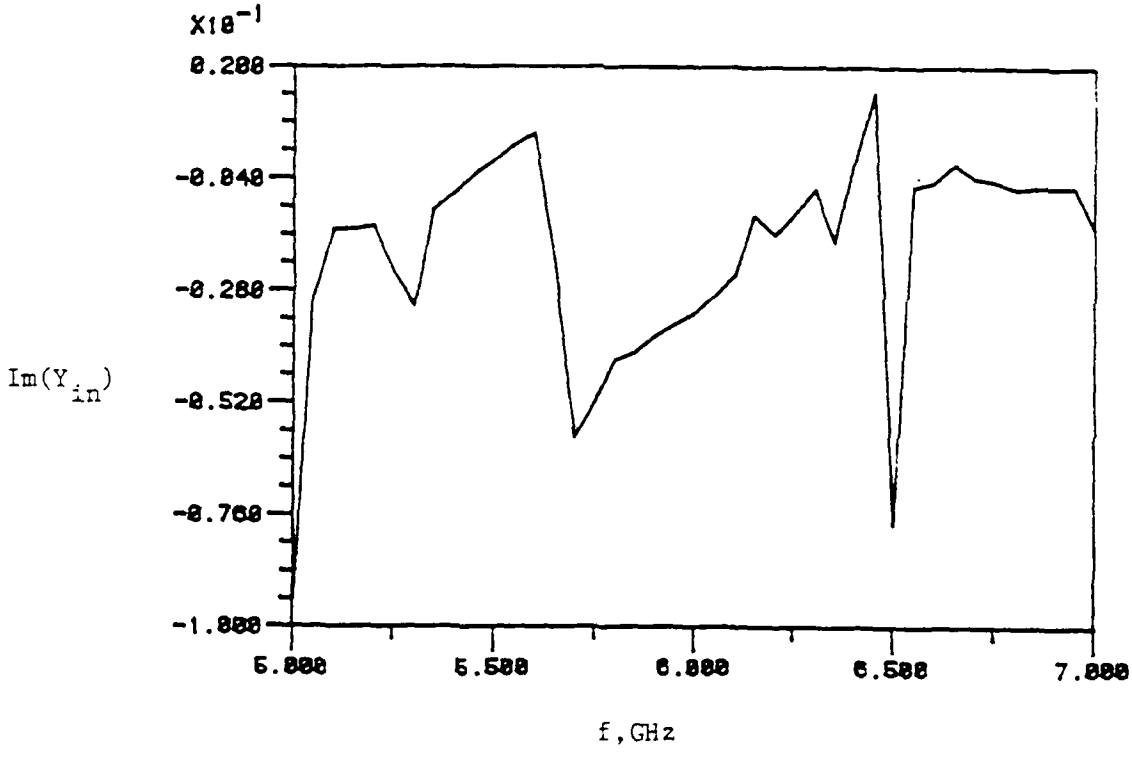
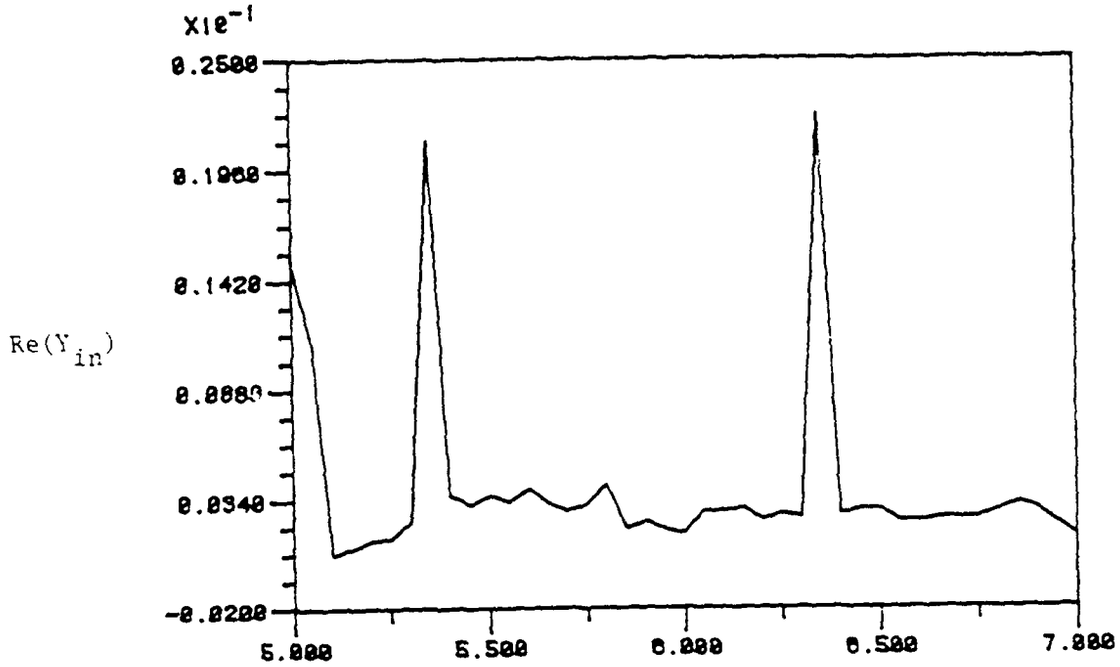


Figure 14. Real and Imaginary Parts of Y_{in} for Board 1.



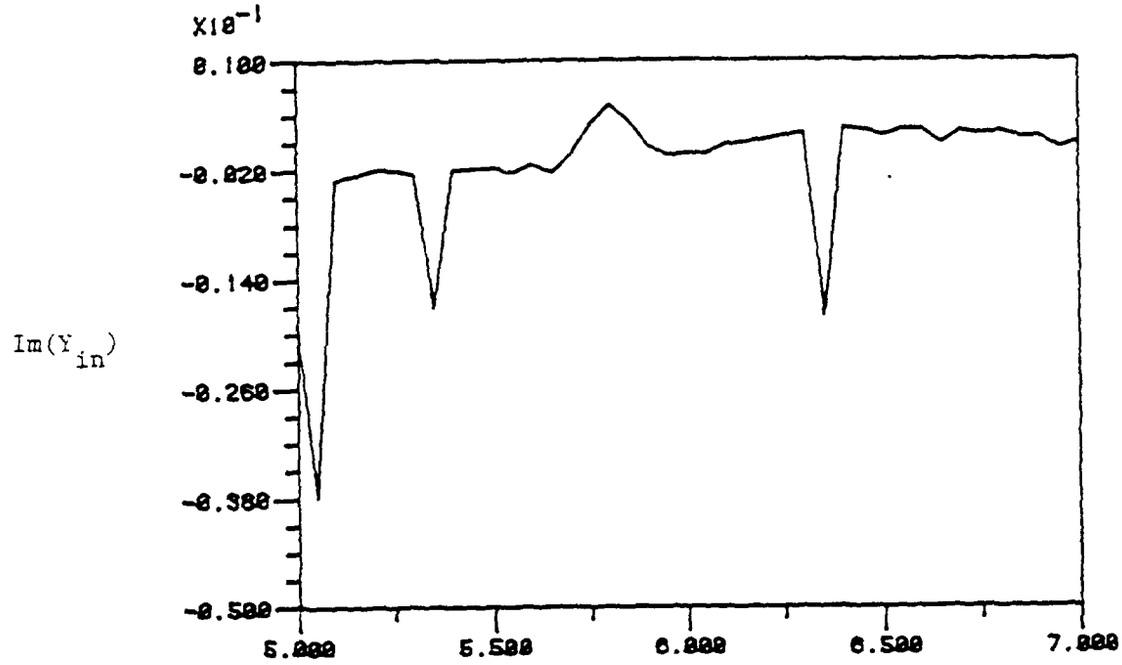


Figure 15. Real and Imaginary Parts of Y_{in} for Board 2.

frequency was 6.0 GHz for all slot antennas so data were taken in the range between 5 GHz and 7 GHz. Figure 14 shows a possible resonance somewhere between 6.3 and 6.5 GHz. Figure 15 shows possible resonances around 5.8 GHz and 6.4 GHz.

To study these phenomena more closely around these frequencies, more data were measured in a narrow band about these frequencies through reflection measurements on the slot antenna boards. These new data were input into the computer program using a linear interpolation scheme on the calibration data to obtain the input admittance over a narrow band. It was observed that there was apparently no resonance at 5.8 GHz. Figures 16 and 17 show the expanded views of Y_{in} between 6.3 and 6.5 GHz for boards 1 and 2, respectively. By looking for zero crossings in the imaginary part and a corresponding peak in the real parts, it was determined that board 1 had a resonance at 6.4 GHz and board 2 had a resonance at 6.36 GHz. These are shown in the Figures. The real parts of Y_{in} are 0.042 S for board 1, and 0.145 S for board 2.

F. Measurement of Far Field Patterns for Slot Radiators Fed by Coplanar Waveguide

Experimental radiation patterns for the slot antennas fed by coplanar waveguide were also measured. The coordinate system relative to the slot is shown in Figure 18. Four E-field patterns were measured. The H-field patterns are proportional to these E-field patterns. Referring to Figure 18, the patterns are:

$$\begin{aligned} &\text{(a)} \quad |E_\theta(\theta)|, \ \phi = 90°, 270° \\ &\text{(b)} \quad |E_\phi(\theta)|, \ \phi = 90°, 270° \\ &\text{(c)} \quad |E_\theta(\theta)|, \ \phi = \ 0°, 180° \\ &\text{(d)} \quad |E_\phi(\theta)|, \ \phi = \ 0°, 180° \end{aligned} \tag{2.15}$$

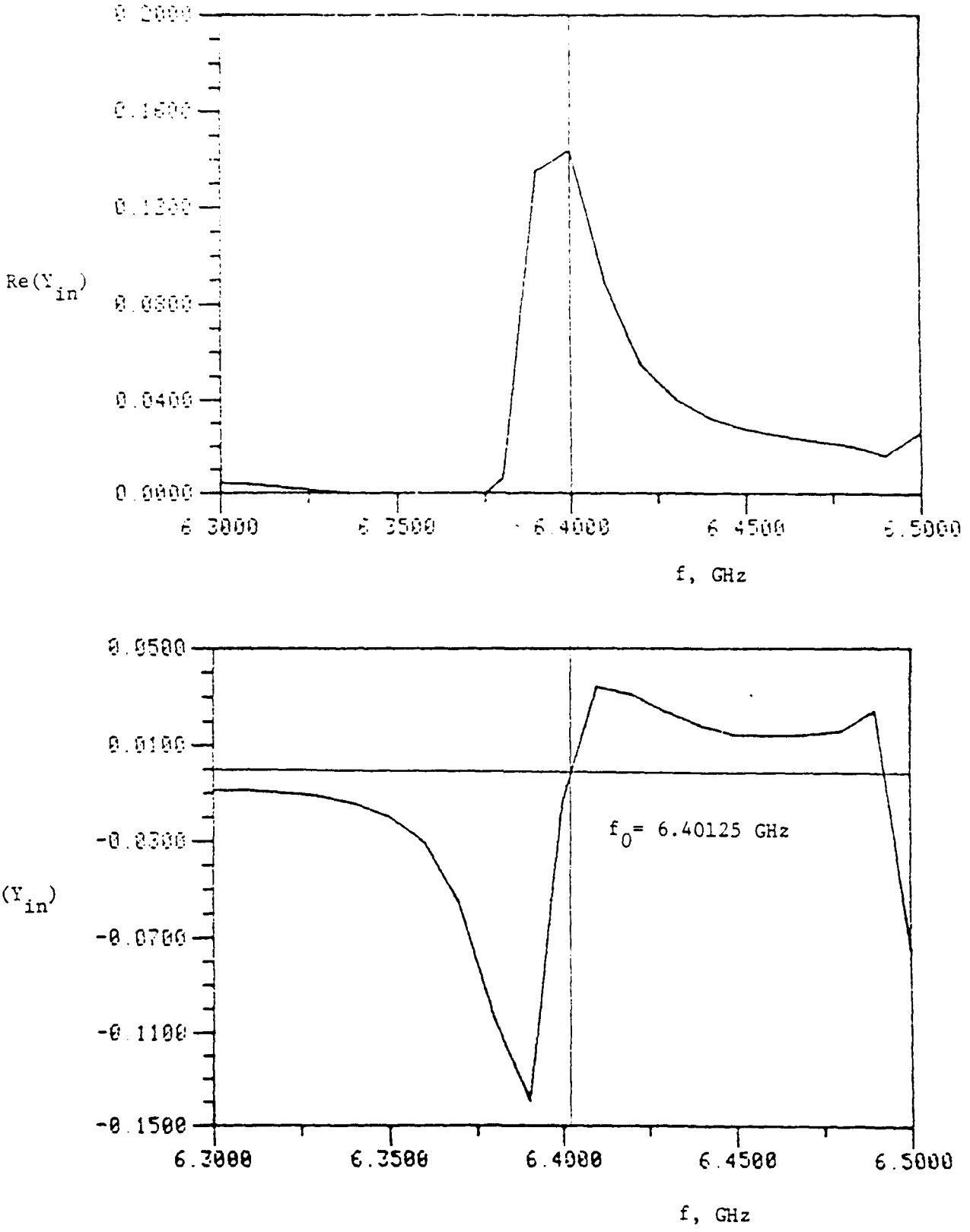


Figure 16. Expanded View with f Between 6.3 and 6.5 GHz of Real and Imaginary Parts of Y_{in} for Board 1.

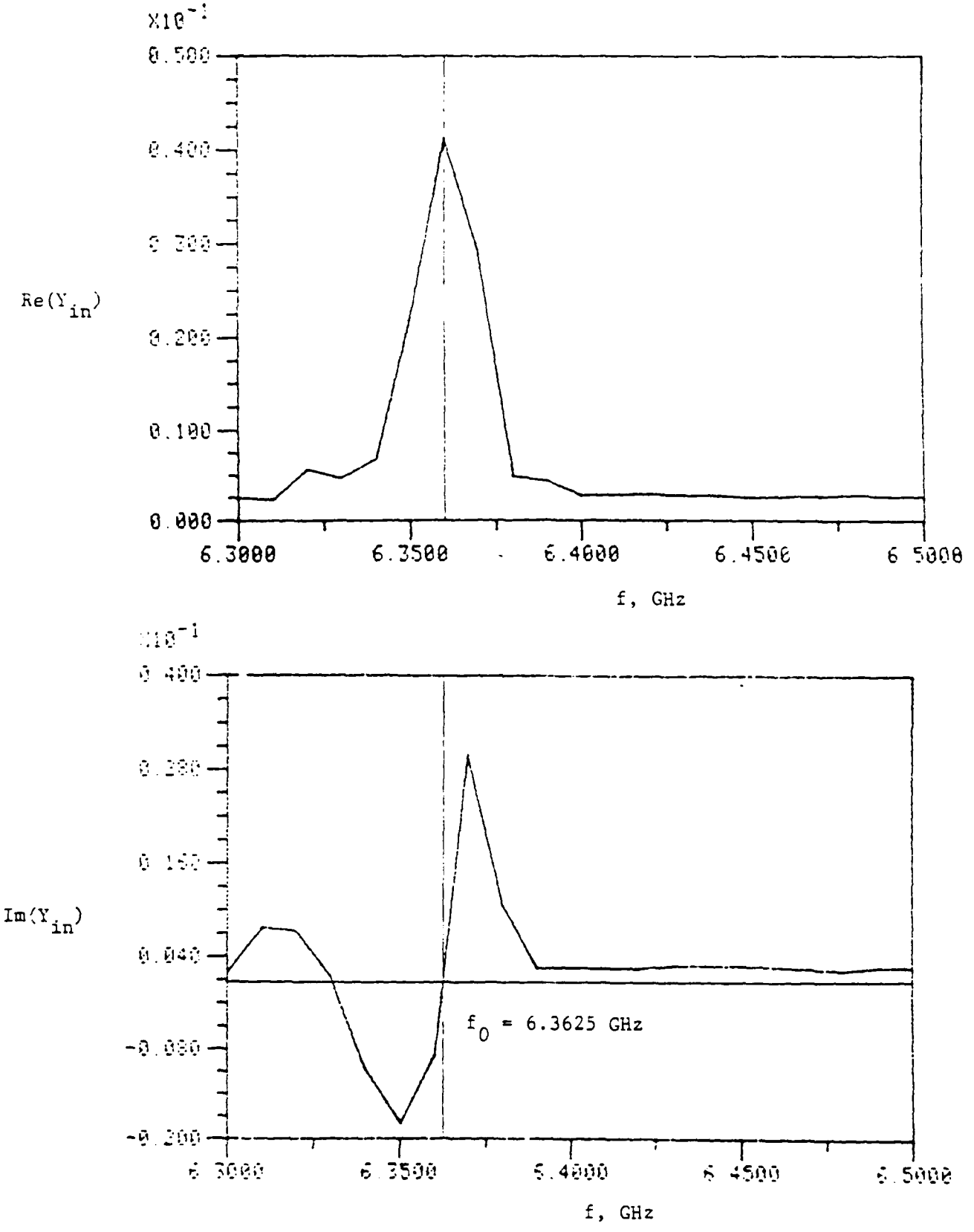


Figure 17. Expanded View with f Between 6.3 and 6.5 GHz of Real and Imaginary Parts of Y_{in} for Board 2.

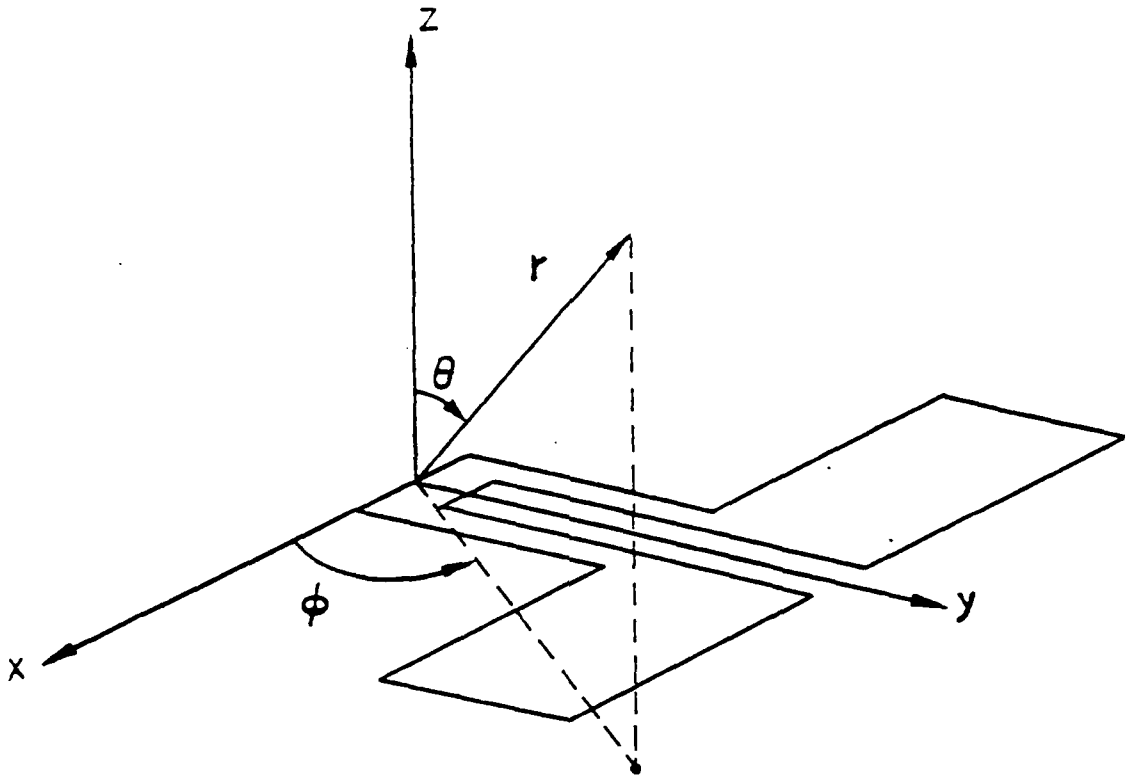


Figure 18. Coordinate System for Patterns.

Cases (a) and (b) are the y-z plane patterns and cases (c) and (d) are the x-z plane patterns. E_θ is in the planes and E_ϕ is perpendicular to the planes. For the y-z plane pattern and slot directed in the x-direction, E_θ is the principle component and E_ϕ is a cross polarized component. For the x-z plane patterns and the slot directed in the x-direction, E_ϕ is the principle component and E_θ is a cross polarized component.

To allow for easy measurement of x-z and y-z plane patterns using a ground plane, the 2m x 2m ground plane that was used has the circular mount shown in Figure 19. The patterns for the x-z plane are measured and then the circular mount is rotated by 90° and the y-z plane patterns are measured.

The measurement set-up is shown in Figure 20. The signal source is a sweep generator used in the CW mode at the resonant frequency of the slot. The generator is modulated using a 1 KHz sine wave generator as shown. The modulated signal is fed to the slot antenna fed by CPW mounted in the ground plane. The ground plane forms one of the perimeter walls of a small anechoic chamber. Inside the chamber, a receiving horn antenna and a crystal detector mount are placed on a supporting stand level with the transmitting antenna. A protractor measures the angle of horn position. The detector output is connected to the VSWR meter (outside the chamber) which indicates a meter reading whose deflection is proportional to received field strength. Funds for a horn and waveguide/coax transistion were provided by the University of Mississippi.

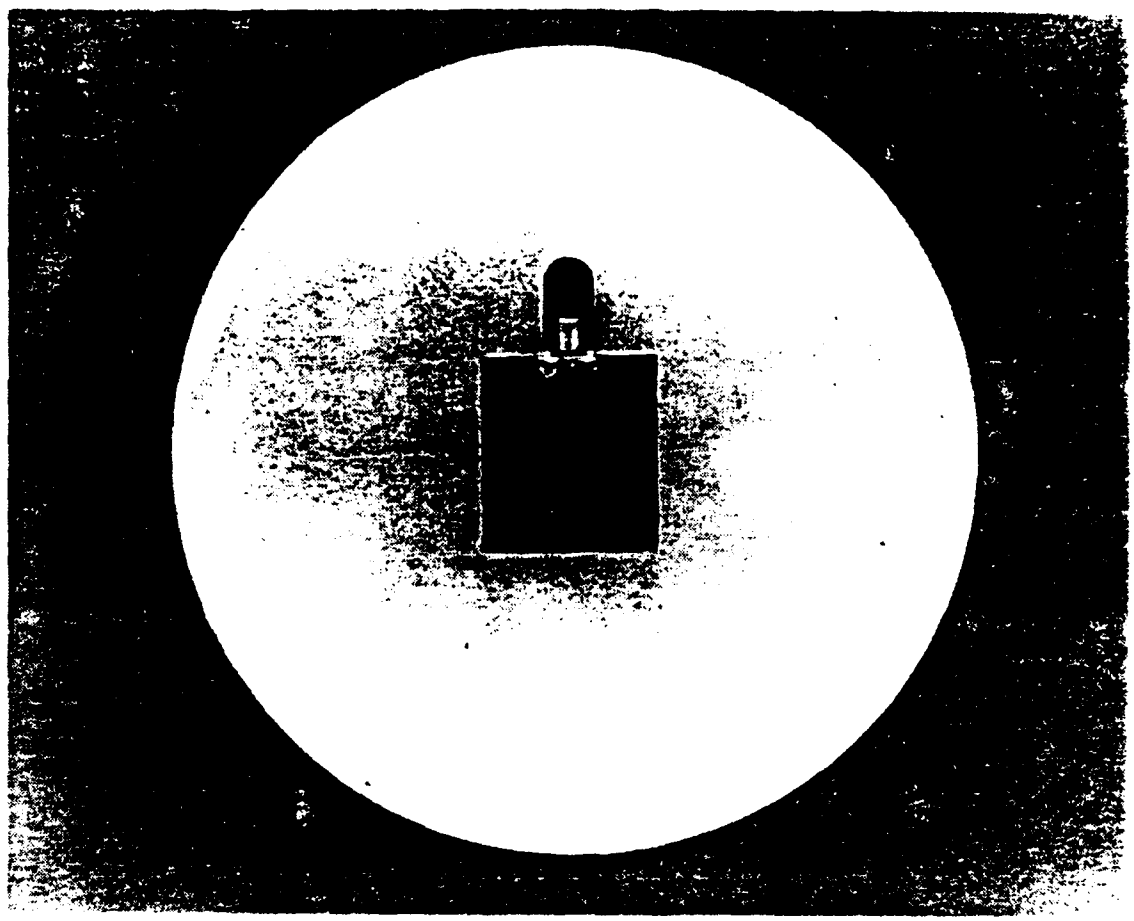


Figure 19. Circular Mount for Ground Plane Mounting of Slot Antenns Fed by CPW.

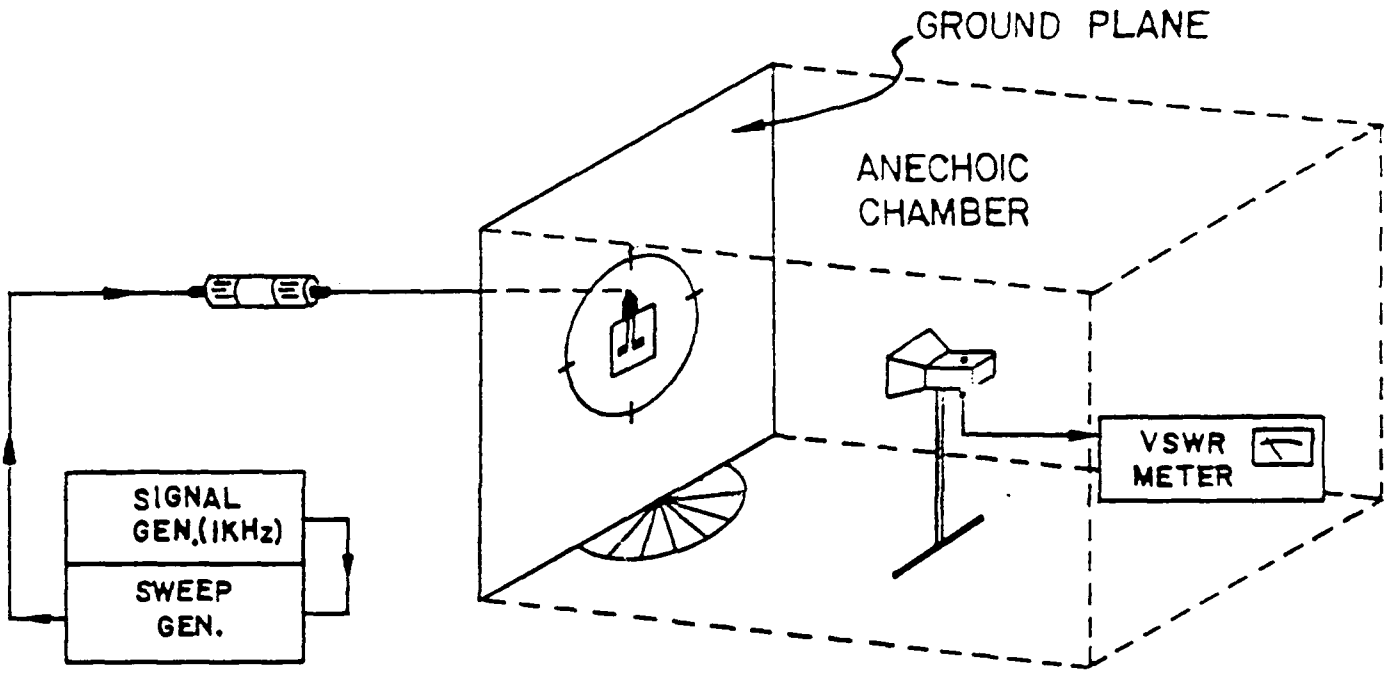


Figure 20. Pattern Measurement Set-up Showing Ground Plane, Anechoic Chamber, and Receiving Horn. A Protractor is also Shown Underneath the Slot Antenna.

The horn has just arrived and so measurements with it have not yet been completed. Preliminary measurements using a horn below cut off at 6.4 GHz have been made. These measurements are quite noisy and need to be re-measured with the new horn. Similar measurements have been performed [20]. The preliminary results for Case 1 of Table 1 are shown in Figure 21(a) through Figure 21(d) for the cases of Equation (2.15). A scale factor between graphs needs to be found. The graphs are quite noisy. The preliminary results for Case 2 of Table 1 are shown in Figure 22(a) through Figure 22(d).

Time didn't permit measuring the antennas of Cases 3 and 4 from Table 1. There were problems in determining the resonant frequency.

G. Conclusions of Measurements

For the antenna of Case 1, the design frequency was 6.0 GHz. The measured resonance occurred at 6.40125 GHz. Therefore, the per cent error is resonant frequency was 6.69%. For the antenna of Case 2, the design frequency was also 6.0 GHz. The measured resonance occurred at 6.3625 GHz. This resulted in a 6.04% error in frequency.

In both cases, the frequency was off by about 6%. The reasons for this need to be checked. Also, more work needs to be done to determine why we had trouble finding the resonance for board 4 (board 3 has not been completed yet). The new horn needs to be used to measure patterns for the antennas.

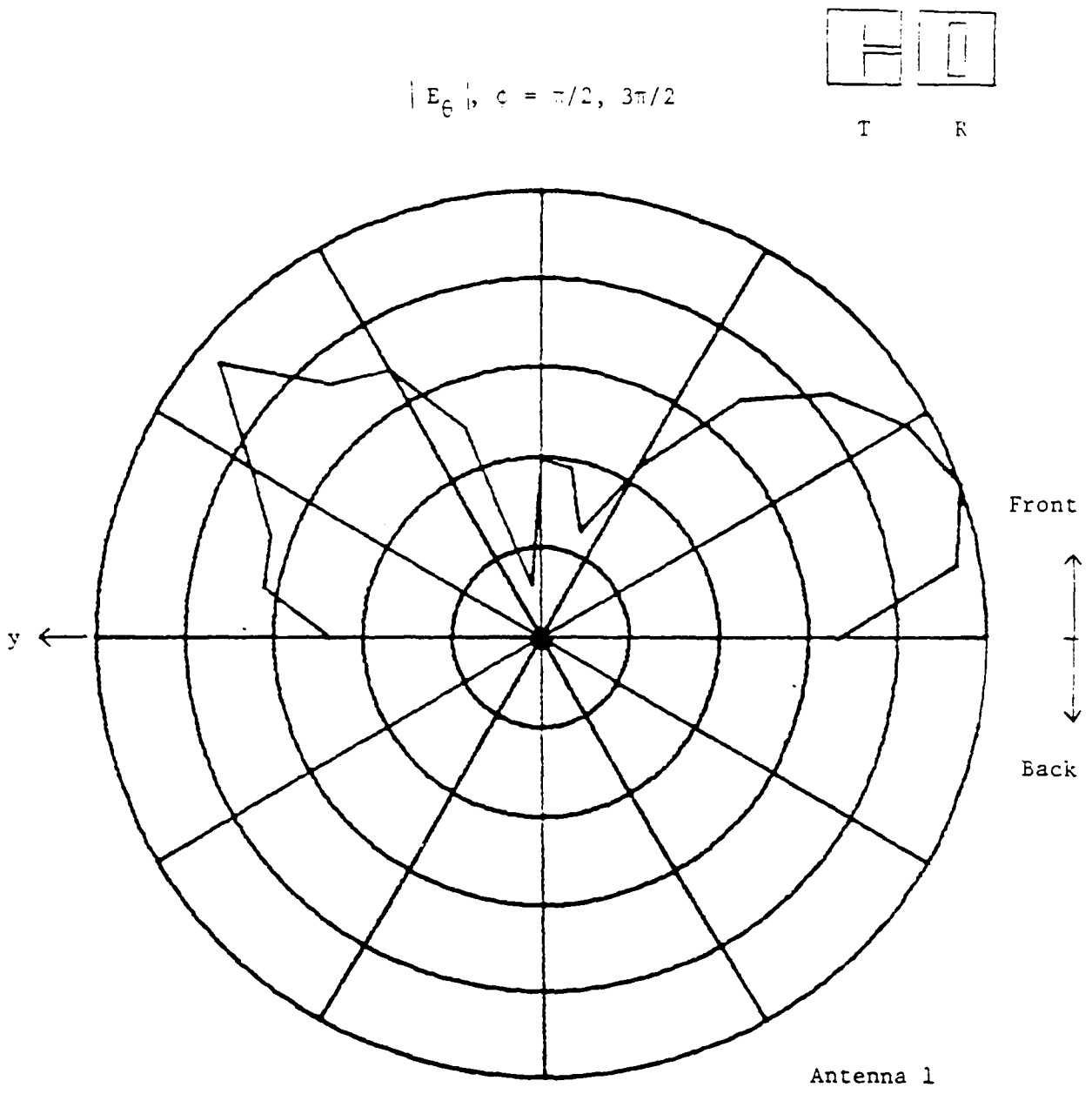


Figure 21(a).

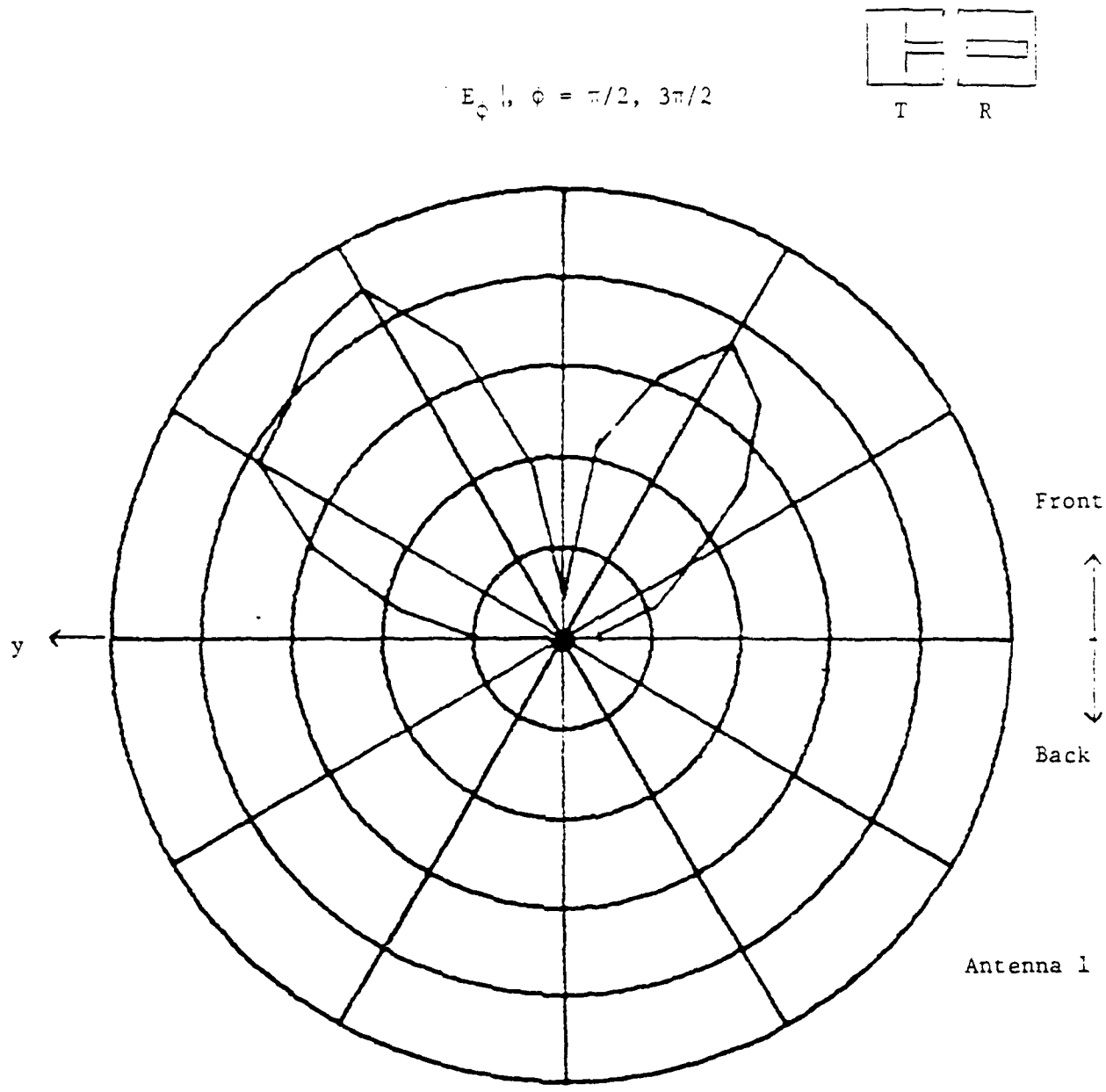


Figure 21(b).

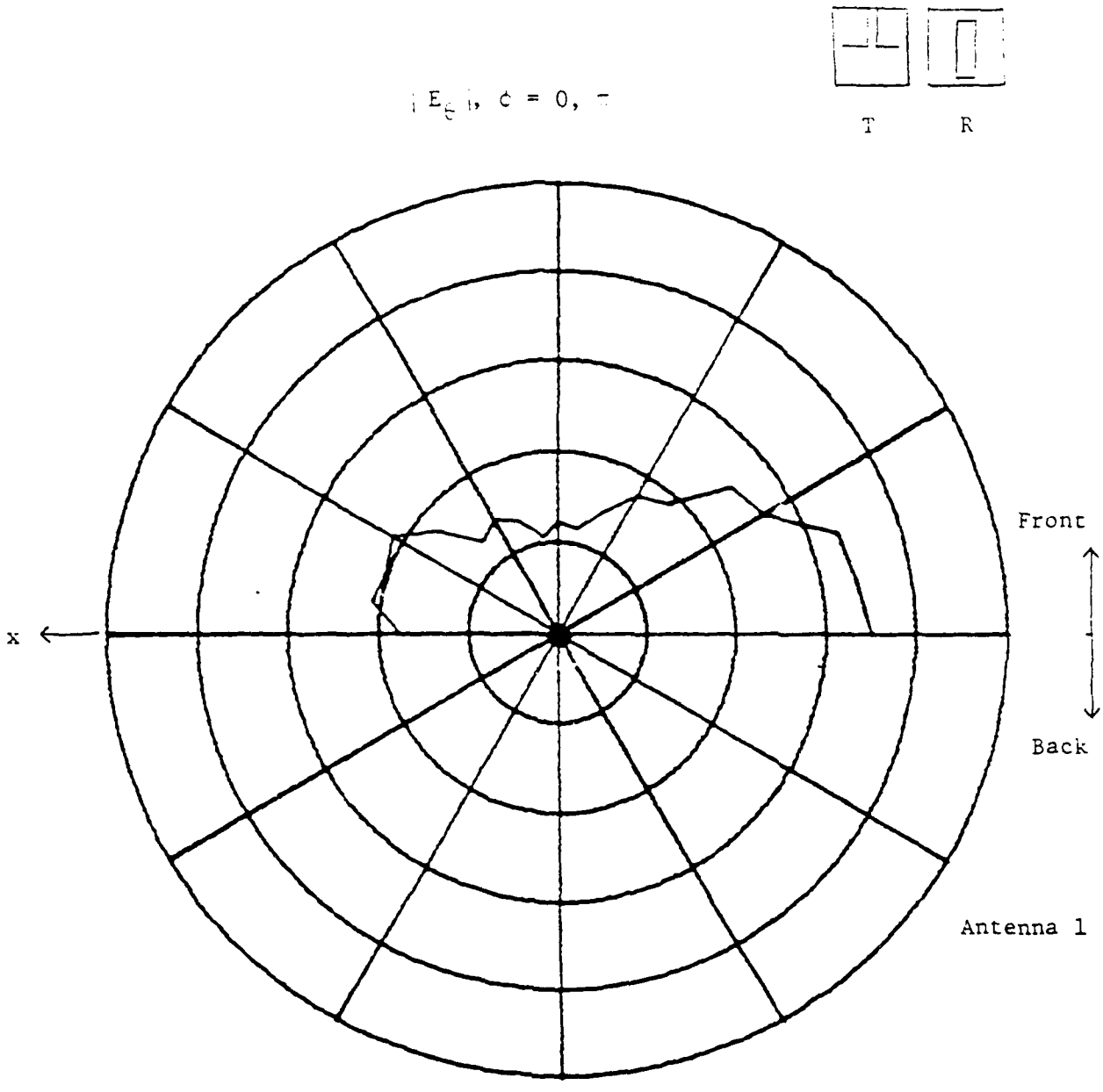


Figure 21(c).

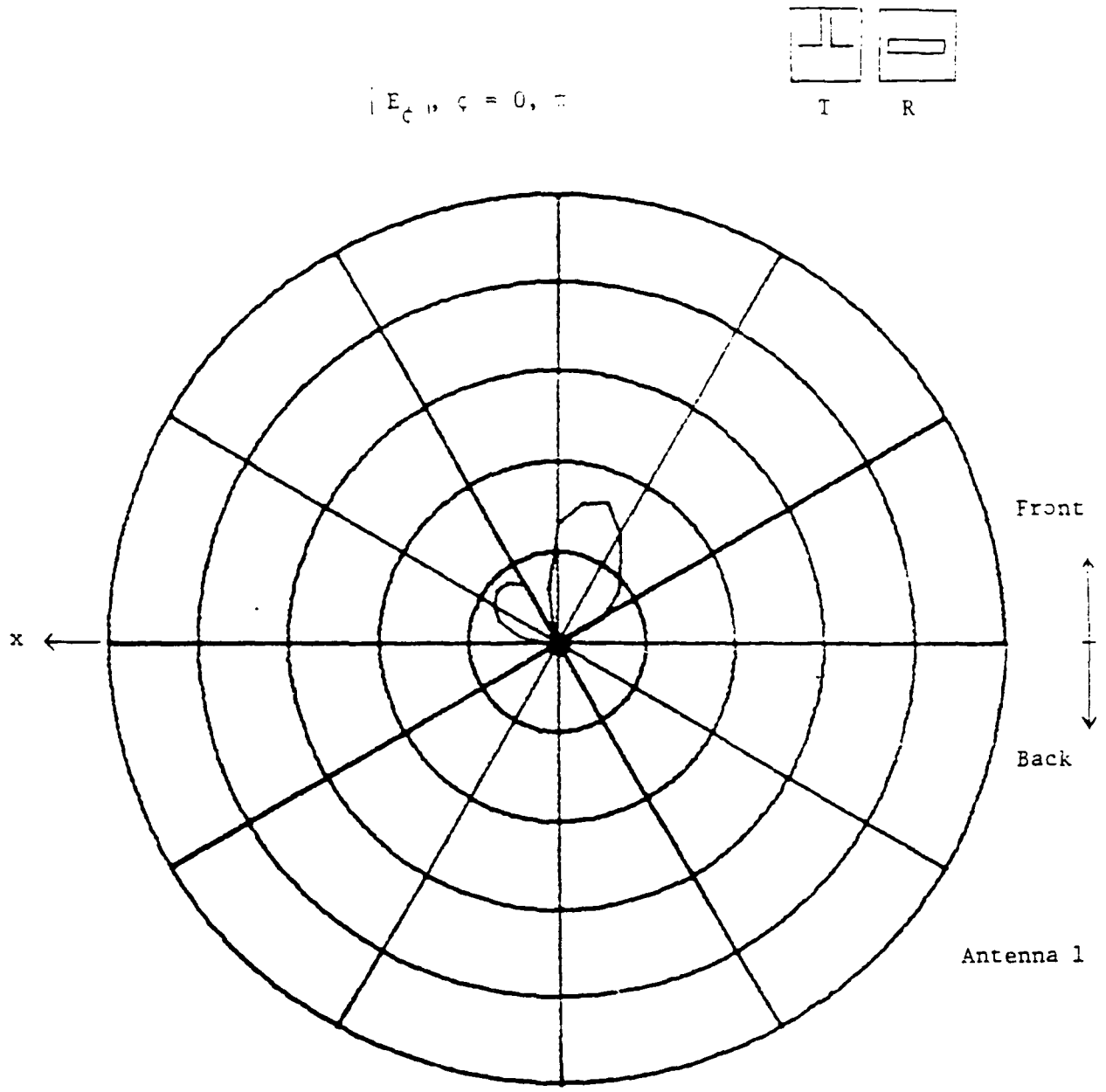
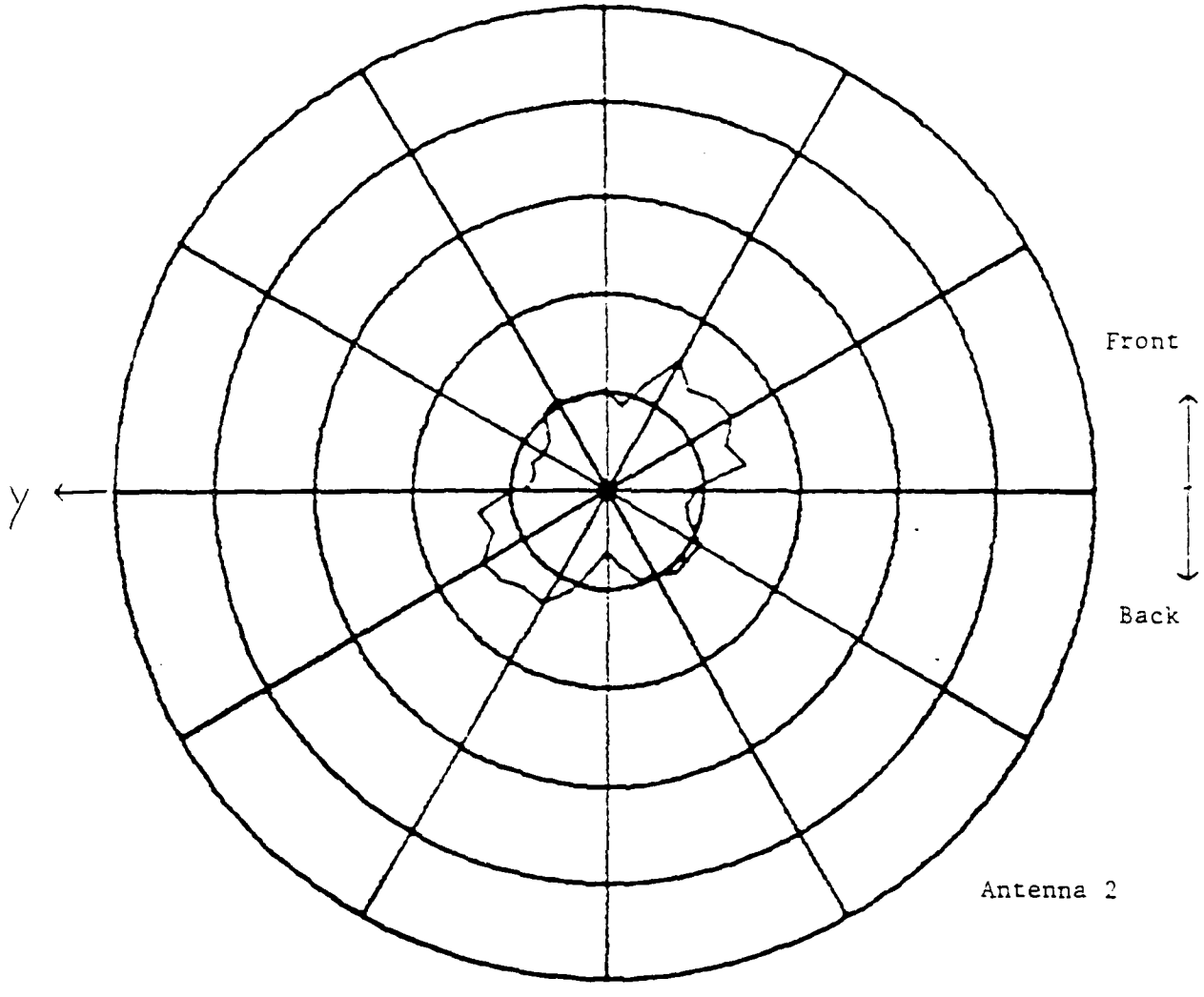


Figure 21(d).

E_θ , $\phi = \pi/2$, $\beta = \pi/2$

T R



Front: Dielectric

Back: Antenna and Feed (CPW)

Figure 22(a).

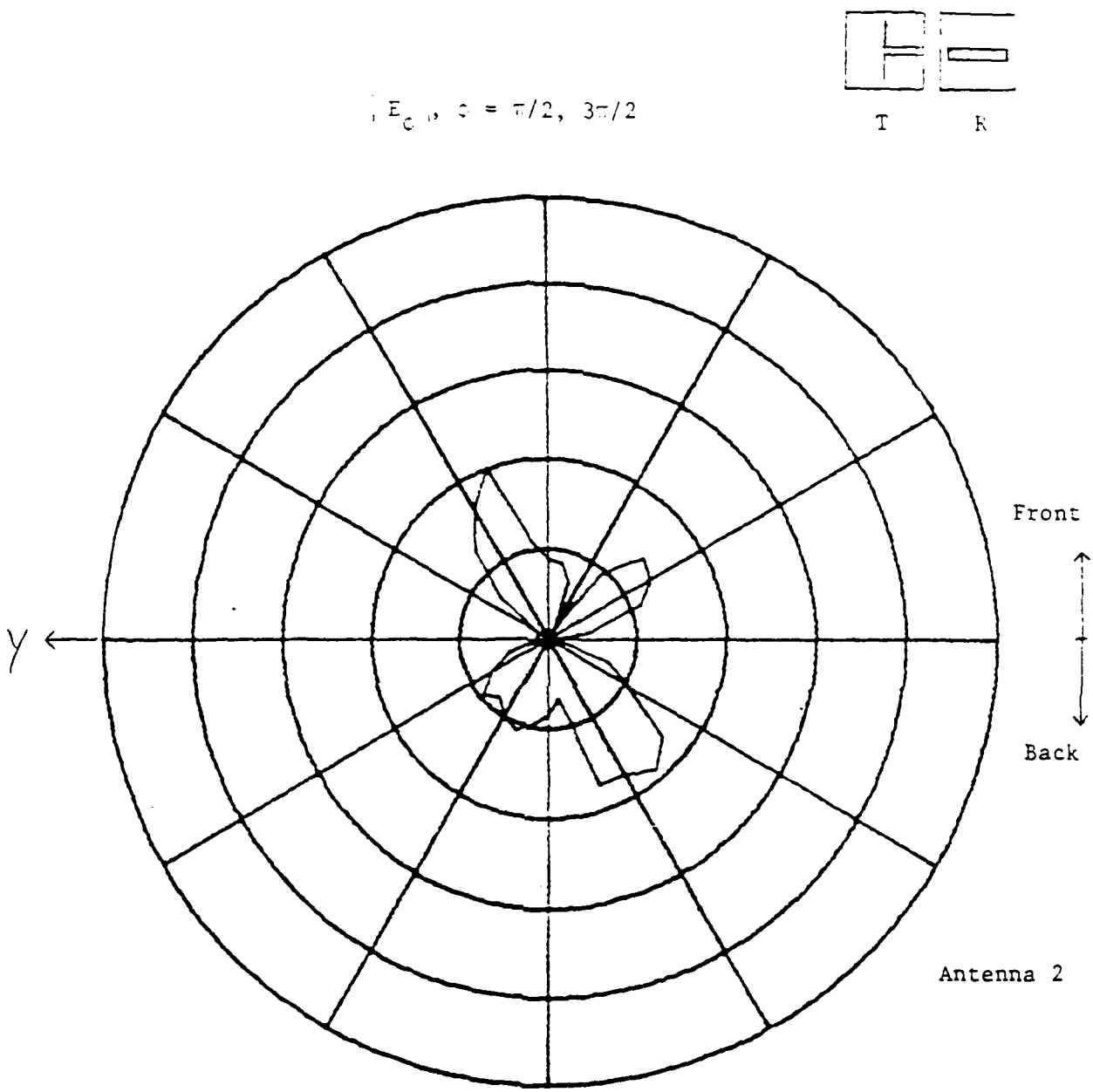


Figure 22(b).

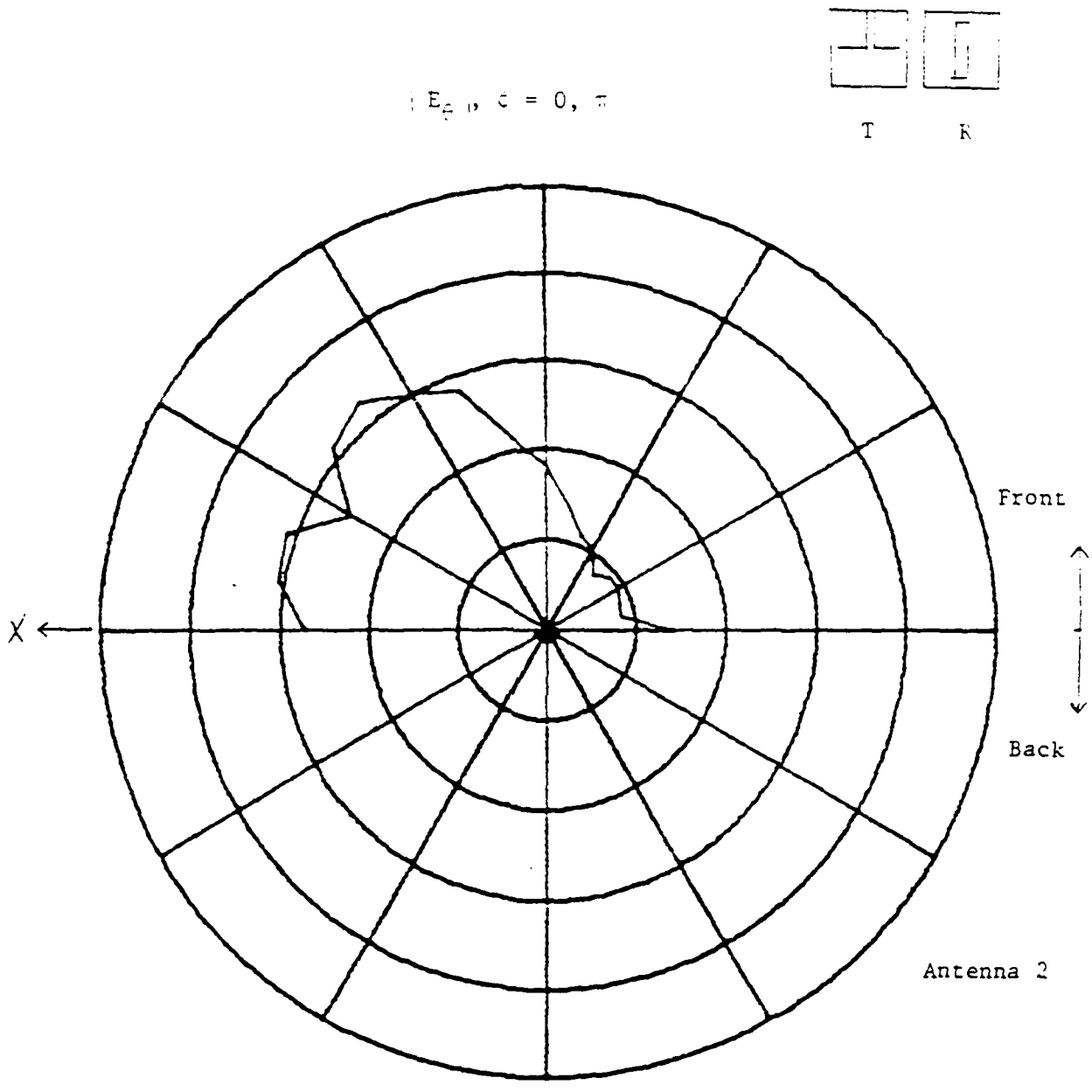


Figure 22(c).

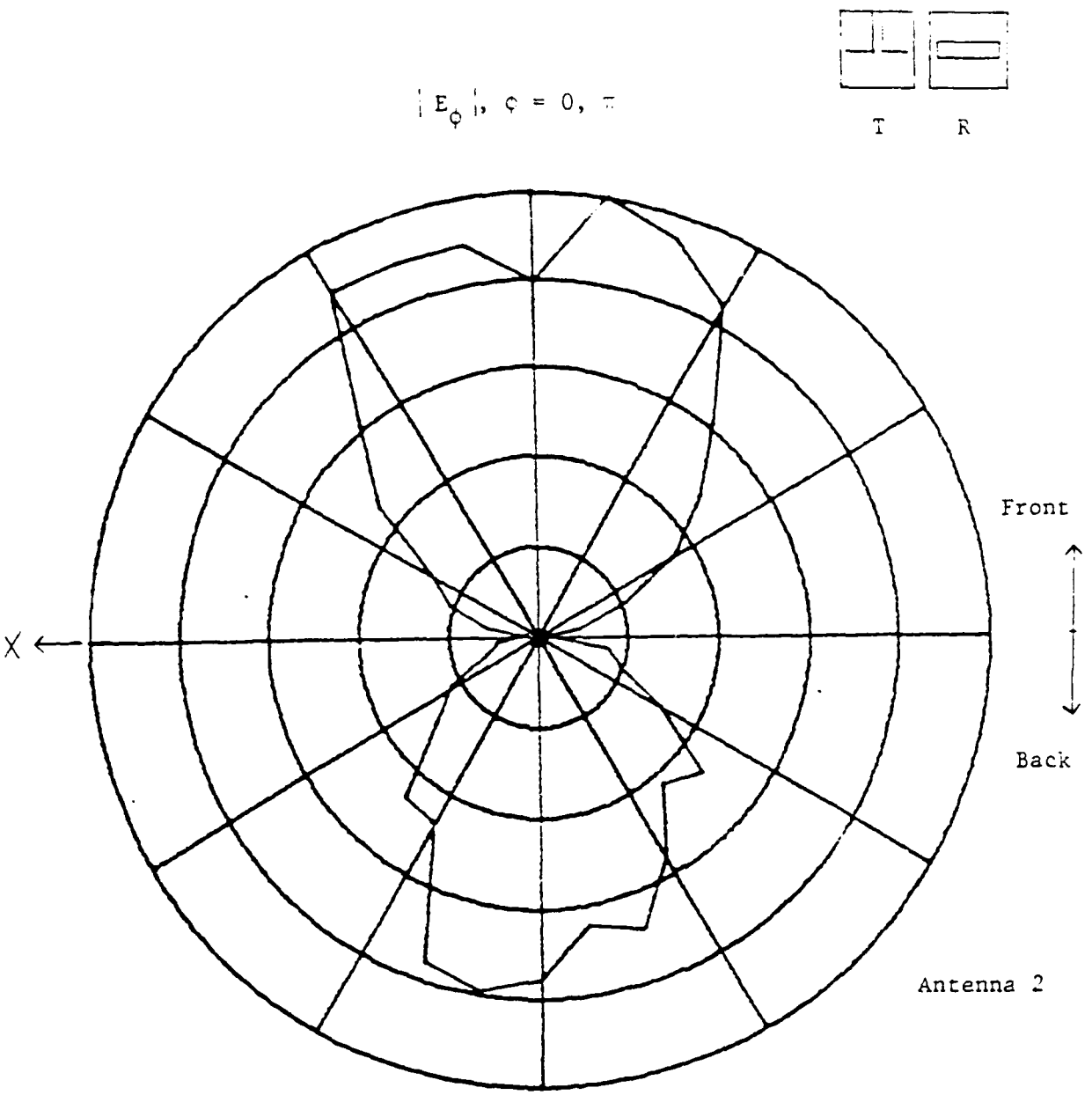


Figure 22(d).

III. MOMENT METHOD FORMULATION OF INFINITE PHASED ARRAYS

In addition to the measurements work just described, a moment method solution for an infinite phased array of slots fed by coplanar waveguide over a dielectric half-space was developed. For the single element case, this writer [1] has provided a moment method solution. Therefore, the present work was undertaken to extend the results to the phased array case.

A. Infinite Phased Array Currents

The infinite array case is straight-forward conceptually. In this case, if one assumes that any element's currents are the same as any other element's currents except perhaps for a progressive phase shift between them and their closest neighbors, then the mathematical formulation can be limited in extent to the study of a single unit cell.

Consider the periodic array shown schematically in Figure 23. Assume the elements are slots in a ground plane and have a progressive phase shift between elements. The x-extent of a cell is D_x and the y-extent is D_y. The cell including the origin is labelled S_D. The field everywhere is unique [21] if the tangential E field is known over the planar surface z = 0. The E field on the plane can be replaced by a magnetic current $\vec{M}_s = \vec{E} \times \hat{z}$ on top of an unslotted plane conductor. This problem has the same solution as the original problem and is an application of the equivalence theorem. The electric field in the slot locations is now the same as it was before and is zero elsewhere.

The magnetic current (or tangential electric field) on the z = 0 surface has the following periodicity property

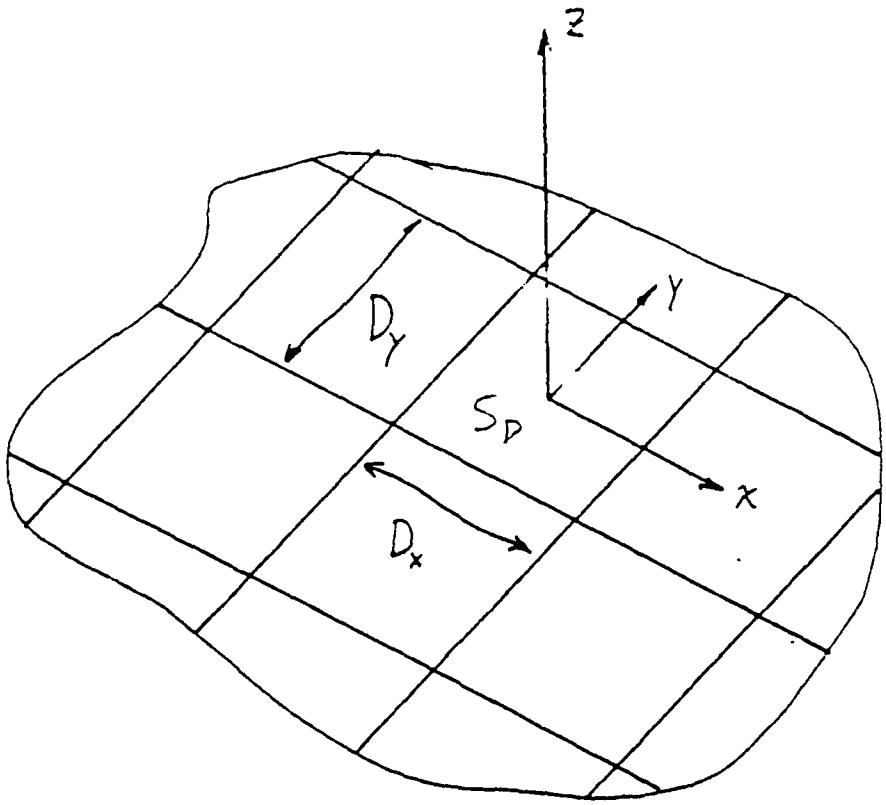


Figure 23. Phased Array.

$$\vec{M}_s\ (x + mD_x,\ y + nD_y) = \vec{M}_s\ (x,y) \left(e^{-jk_0u_0D_x}\right)^m \left(e^{-jk_0v_0D_y}\right)^n \quad (3.1)$$

$$-\frac{D_x}{2} < x < \frac{D_x}{2}$$

m,n integers

$$-\frac{D_y}{2} < y < \frac{D_y}{2}$$

where [22], $u_0 = \sin\theta_0 \cos\phi_0$, $v_0 = \sin\theta_0 \sin\phi_0$. The complex exponentials account for the progressive phase shift mentioned previously and are allowed in Floquet's Theorem for solving periodic differential equations. Floquet's Theorem is formally described in several books [23-25]. Multiplying the above equation through by $\exp(jk_0u_0\ (x + mD_x) + jk_0v_0\ (y + nD_y))$, one obtains [26]

$$\vec{M}_s\ (x,y)\ e^{jk_0u_0x}\ e^{jk_0v_0y} =$$

$$= e^{jk_0u_0\ (x + mD_x)}\ e^{jk_0v_0\ (y + nD_y)}\ \vec{M}_s\ (x + mD_x,\ y + nD_y) \quad (3.2)$$

This is now written in a proper form to see the periodic nature of the function. $M_s(x,y)$ itself is not periodic.

This periodic function is now in a form suitable for expansion in a complex exponential Fourier series

$$\vec{M}_s\ (x,y)\ e^{jk_0u_0x}\ e^{jk_0v_0y} =$$

$$= \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \vec{A}_{pq}\ e^{-jk_0p\frac{x}{D_x}}\ e^{-jk_0q\frac{v}{D_y}} \quad (3.3)$$

The current M_s can then be written in the form

$$\vec{M}_s(x,y) = \tag{3.4}$$

$$= e^{-jk_0u_0x}\, e^{-jk_0v_0y} \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \vec{A}_{pq}\, e^{-jk_0p\frac{x}{D_x}}\, e^{-jk_0q\frac{y}{D_y}}$$

This is the form given by Floquet's Theorem [23]. This will not be used here, but instead a moment method integral equation solution will be found.

Assuming an $e^{-j\Gamma z}$ variation with z, the electric and magnetic fields due to the magnetic current can likewise be expanded in a Floquet series [25]. This can be viewed as the field in a waveguide due to a current in the guide. The outward propagation constant is given by

$$\Gamma_{pq}^2 = k_0^2 - \left(k_0u_0 + k_0\frac{p\lambda}{D_x}\right)^2 - \left(k_0v_0 + k_0\frac{q\lambda}{D_y}\right)^2 \tag{3.5}$$

Each (p,q) represents a single Floquet mode. Green's functions for planar slots in phased arrays are given by Mailloux [22].

B. Integral Equation Formulation

The integral equation is derived by shorting the slots and covering the slots with equivalent magnetic currents $\vec{M}_s$ on both sides of the conducting sheet in the z=0 plane. Applying image theory effectively doubles the magnetic currents which now reside in homogeneous space. These currents are $\vec{M}_1 = -2\,\vec{M}_s$ and $\vec{M}_0 = 2\,\vec{M}_s$. Regions 0 and 1 are identified as free space and dielectric, respectively. One obtains

$$\vec{H}_0^s = -j\omega\vec{F}_0 - \nabla\psi_0 \tag{3.6}$$

and

$$\vec{H}_1^s = -j\omega\vec{F}_1 - \nabla\psi_1$$

where $\vec{F}$ and ψ are the vector and scalar potentials, respectively.

The currents $\vec{M}_0(x',y')$ and $\vec{M}_1(x',y')$ each consist of a planar group

of sources. These currents can be written as $\vec{M}_a(x',y')$, a=0,1, for x' and y' in $(-\infty,\infty)$. The x' and y' values can be written as

$$\begin{aligned} x' &= x'_0 + mD_x \qquad & x' &\in (-\infty,\infty) \\ & & x'_0 &\in (-D_x/2, D_x/2) \\ & & m &\in (-\infty,\infty) \\ y' &= y'_0 + nD_y \qquad & y' &\in (-\infty,\infty) \\ & & y'_0 &\in (-D_y/2, D_y/2) \\ & & n &\in (-\infty,\infty) \end{aligned} \tag{3.7}$$

The currents can also be written as

$$\vec{M}_a(x',y') = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} M_a(x'_0 + mD_x, y'_0 + nD_y), \qquad a=0,1 \tag{3.8}$$

The electric vector potential can be written as

$$\vec{F}_a(x,y,z) = \epsilon_a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \vec{M}_a(x',y') \frac{e^{-jk_a R}}{4\pi R} dx'dy' \tag{3.9}$$

where $R = \sqrt{(x-x')^2 + (y-y')^2 + z^2}$. Substituting Eq. (3.7) and Eq. (3.8) in Eq. (3.9), one obtains

$$\vec{F}_a(x,y,z) = \epsilon_a \int_{S_D} \int \sum_m \sum_n \vec{M}_a (x'_0 + mD_x, y'_0 + nD_y) \frac{e^{-jk_a R_{mn}}}{4\pi R_{mn}} dx'_0 dy'_0 \tag{3.10}$$

where $R_{mn} = \sqrt{(x'_0+mD_x - x)^2 + (y'_0+nD_y - y)^2 + z^2}$ and S_D is the cell including the origin (when m=n=0). From Eq. (3.1), one obtains the periodic nature of the current $\vec{M}_a$, a=0.1. Therefore,

$$\vec{M}_a(x'_0+mD_x, y'_0+nD_y) = \vec{M}_a(x'_0, y'_0)\, e^{-jk_a u_0 mD_x}\, e^{-jk_a v_0 nD_y}, \qquad a=0,1 \tag{3.11}$$

Substituting Eq. (3.11) in Eq. (3.10), one obtains

$$\vec{F}_a(x,y,z) = \epsilon \int_{S_D}\int \sum_m \sum_n \vec{M}_a\,(x_0',y_0')\, e^{-jk_a u_0 m D_x}\, e^{-jk_a v_0 n D_y}\, \frac{e^{-jk_a R_{mn}}}{4\pi R_{mn}}\, dx_0' dy_0' \tag{3.12}$$

$$= \epsilon_a \int_{S_D}\int \vec{M}_a(x_0',y_0')\, G(k_a;x-x_0',y-y_0',z)\, dx_0' dy_0'$$

where

$$G(k_a;x-x_0',y-y_0',z) = \sum_m \sum_n e^{-jk_a u_0 m L_x}\, e^{-jk_a v_0 n D_y}\, \frac{e^{-jk_a R_{mn}}}{4\pi R_{mn}}, \qquad a=0,1 \tag{3.13}$$

Therefore, for the free space case (a=0), one can write

$$\vec{F}_0(x,y,z) = \epsilon_0 \int_{S_D}\int \vec{M}_0\,(x_0',y_0')\, G(k_0;x-x_0',y-y_0',z)\, dx_0'\, dy_0'$$

$$\psi_0(x,y,z) = \frac{1}{\mu_0} \int_{S_D}\int m_0\,(x_0',y_0')\, G(k_0;x-x_0',y-y_0',z)\, dx_0'\, dy_0' \tag{3.14}$$

where $m_0 = \frac{-1}{j\omega} \nabla \cdot \vec{M}_0$

Similarly, for the dielectric case (a=1), one obtains

$$\vec{F}_1(x,y,z) = \epsilon_1 \int_{S_D}\int \vec{M}_1\,(x_0',y_0')\, G(k_1;x-x_0',y-y_0',z)\, dx_0'\, dy_0'$$

$$\psi_1(x,y,z) = \frac{1}{\mu_1} \int_{S_D}\int m_1\,(x_0',y_0')\, G(k_1;x-x_0',y-y_0',z)\, dx_0'\, dy_0' \tag{3.15}$$

where $m_1 = \frac{-1}{j\omega} \nabla \cdot \vec{M}_1$.

Assuming H_1^i (incident) = 0, one finds

$$H_0^t = H_0^s + H_0^i$$

$$H_1^t = H_1^s \tag{3.16}$$

The integral equation can be found by enforcing tangential $\vec{H}^t$ (total) to be continuous through the slots

$$\lim_{z\uparrow 0} \hat{z}\times H_0^t = \lim_{z\downarrow 0} \hat{z}\times H_1^t \qquad \text{through } S_a \text{ (aperture)} \tag{3.17}$$

This results in the coupled integral equation

$$j\omega\epsilon_1 F_{s1y}(x,y,0) + j\omega\epsilon_0 F_{s0y}(x,y,0) - \frac{1}{j\omega\mu_1}\partial\psi_{s1}(x,y,0)/\partial y - \frac{1}{j\omega\mu_0}\partial\psi_{s0}(x,y,0)/\partial y = H_y^{sci}(x,y,0)/2 \text{ in slot} \tag{3.18a}$$

$$j\omega\epsilon_1 F_{s1x}(x,y,0) + j\omega\epsilon_0 F_{s0x}(x,y,0) - \frac{1}{j\omega\mu_1}\partial\psi_{s1}(x,y,0)/\partial x - \frac{1}{j\omega\mu_0}\partial\psi_{s0}(x,y,0)/\partial x = H_x^{sci}(x,y,0)/2 \text{ in slot} \tag{3.18b}$$

where

$$\vec{F}_{s1}(x,y,z) = \int\int_{S_D} \vec{M}_s(x_0',y_0')\, G(k_1;x-x_0',y-y_0',z)\, dx_0'\, dy_0'$$

$$\vec{F}_{s0}(x,y,z) = \int\int_{S_D} \vec{M}_s(x_0',y_0')\, G(k_0;x-x_0',y-y_0',z)\, dx_0'\, dy_0' \tag{3.19}$$

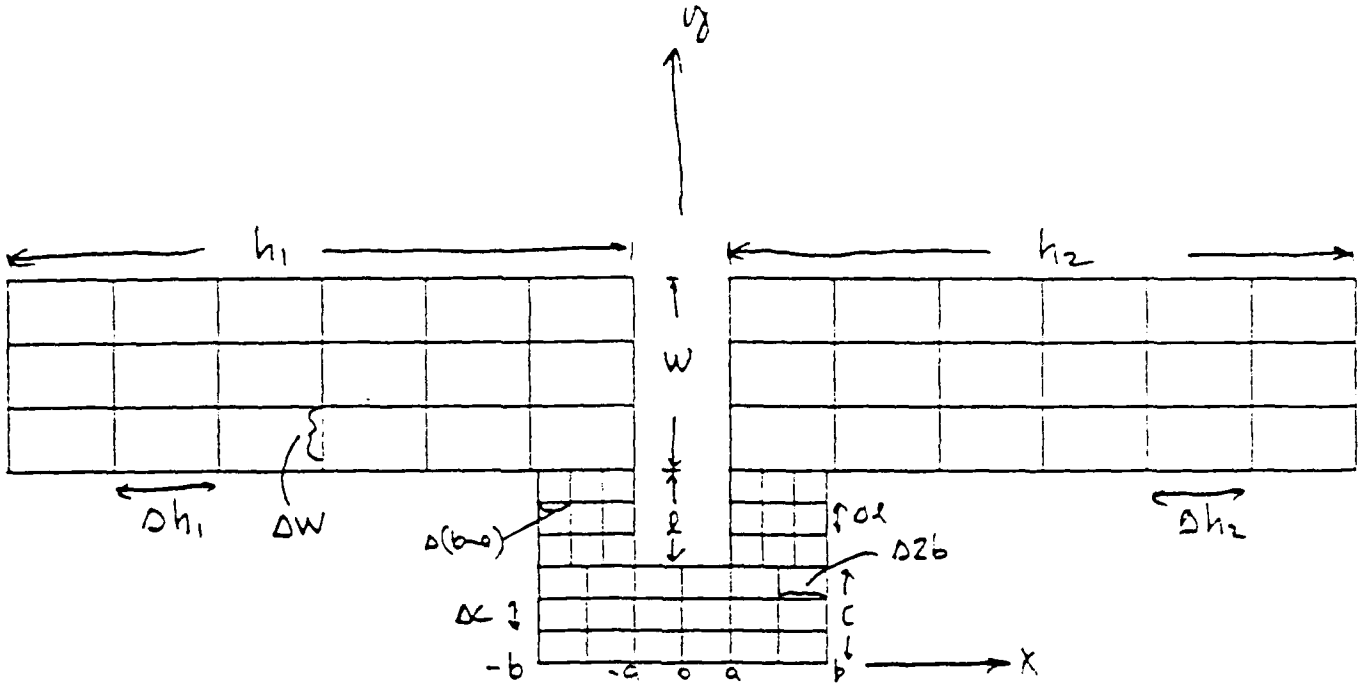
$$\psi_{s1}(x,y,z) = \int\int_{S_D} \nabla\cdot M_s(x_0',y_0')\, G(k_1;x-x_0',y-y_0',z)\, dx_0'\, dy_0'$$

$$\psi_{s0}(x,y,z) = \int\int_{S_D} \nabla\cdot M_s(x_0',y_0')\, G(k ;x-x_0',y-y_0',z)\, dx_0'\, dy_0' \tag{3.20}$$

and $G(k_a;x-x_0',y-y_0',z)$ is given in Eq. (3.13). Eq. (3.18) is the integral equation that needs to be solved.

C. The Moment Method Solution

Figure 24 shows the dimensions of a slot antenna fed by CPW. Unknowns were taken in both the possible magnetic current directions, x



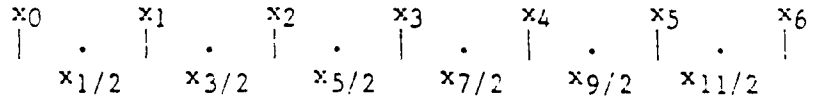


Figure 24. Dimensions of Slot Antenna.

and y. Both current components were obtained [27]. Three integers in y, N_1, N_2, and N_3 and four integers in x, M_1, M_2, M_3, and M_4, partially describe the expansion functions. The Δ's shown in Figure 24 are given by

$$\begin{aligned}
\Delta w &= w/(N_1 + 1) \\
\Delta \ell &= \ell/(N_2 + 1) \\
\Delta c &= c/(N_3 + 1) \\
\\
\Delta h_1 &= h_1/(M_1 + 1) \\
\Delta h_2 &= h_2/(M_2 + 1) \\
\Delta(b-a) &= (b-a)/(M_3 + 1) \\
\Delta 2b &= 2b/(M_4 + 1)
\end{aligned} \tag{3.21}$$

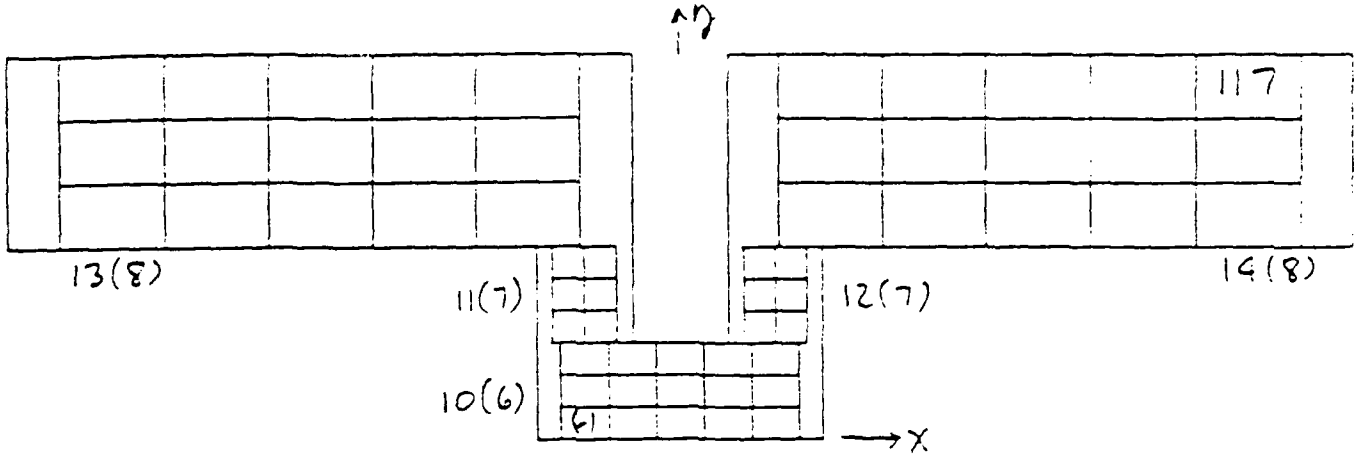
Note that this allows for off-center feeds.

Different expansion domains are taken for the x and y cases. This is because the magnetic current $\vec{M}_s$ is zero for $\vec{M}_s$ normal to the edge. For $\vec{M}_s$ parallel to the edge, the magnetic current is singular as $s^{-1/2}$ away from the edge. The expansion domains are shown in Figure 25.

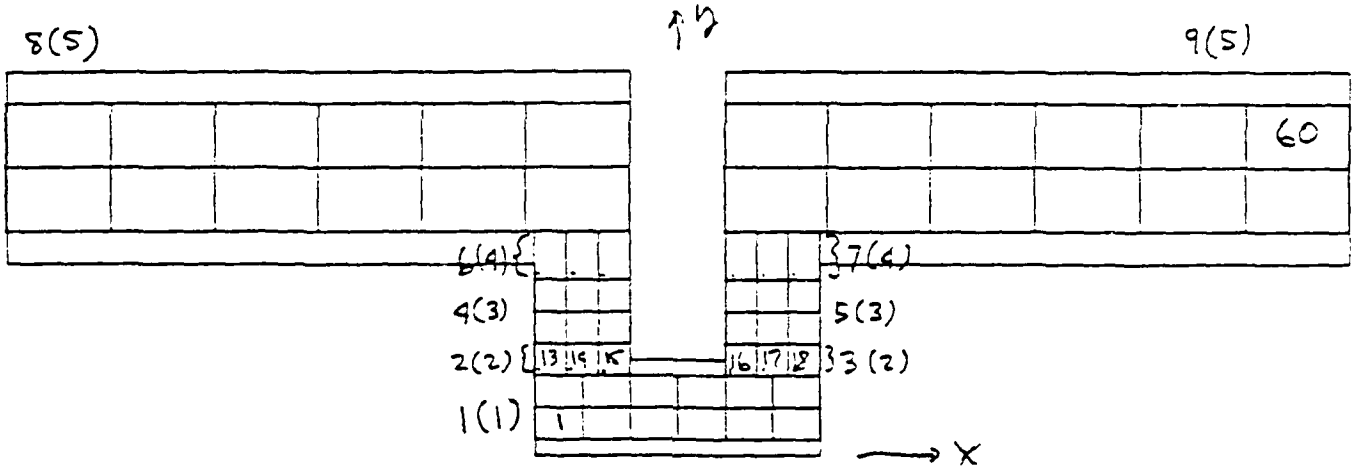
There are three ways of numbering the domains. First, there is continuous numbering. For the example shown, this is from 1 to 60 for the y case and from 61 to 117 for the x case. This numbering starts with the y case and proceeds from left to right. The five individual larger rectangular regions are also numbered together with the four regions present in the y case to insure continuity of the magnetic current. These are numbered in two ways. First, from 1 to 14 and second, from (1) to (8). The first numbers are shown and the second are given in parenthesis in Figure 25. For example, regions 2 and 3 are contained in (2). For regions 1, 4, 5, 8, and 9 in the y case,

$$\begin{aligned}
n &= 1, 2, 3, \ldots, N && \text{in y} \\
m &= 1, 2, 3, \ldots, M+1 && \text{in x}
\end{aligned} \tag{3.22}$$

For regions 10, 11, 12, 13, and 14 in the x case



x-directed expansion domains



y-directed expansion domains

Figure 25. The Magnetic Current Expansion Domains for the General Case.

$$\begin{aligned} n &= 1, 2, 3, \ldots, N+1 \qquad \text{in } y \\ m &= 1, 2, 3, \ldots, M \qquad \text{in } x \end{aligned} \tag{3.23}$$

This numbering allows for a half pulse of zero value on the edges where the magnetic current is normal to the edge. Therefore, any particular square can be identified by [region#, m, n]. Table 4 shows this in detail. Several mapping functions are used in the computer program. These are given in Figure 26. MAPBGN maps from parenthesized numbers to ordinary numbers and MUNSEG is the inverse mapping. MAP12 gives the number of rectangles (1 or 2) in each parenthesized number zone. MAPOFF is useful to convert a 10 through 14 series to an equivalent 1 through 9 number.

Regions 2, 3, 6, and 7 are special regions present for continuity of the magnetic current. For these four regions, the integer M_5 is input so that

$$\Delta_5 = (b-a)/(M_5+1) \tag{3.24}$$

Therefore, m goes from 1 to M_5+1 while n is always 1 for regions 2, 3, 6, and 7.

The current was expanded in the following manner:

$$\begin{aligned} M_y = &\sum_{n=1}^{N_3} \sum_{m=1}^{M_4+1} M^1_{mn} \Lambda^1_{mn} \\ &+ \sum_{m=1}^{M_5+1} (M^2_{m1} \Lambda^2_{m1} + M^3_{m1} \Lambda^3_{m1}) \\ &+ \sum_{n=1}^{N_2} \sum_{m=1}^{M_3+1} (M^4_{mn} \Lambda^4_{mn}(r) + M^5_{mn} \Lambda^5_{mn}(r)) \end{aligned} \tag{3.26a}$$

Table 4. Unknowns in Each Region

Region	X	by	Y
1	M_4+1		N_3
2	M_5+1		1
3	M_5+1		1
4	M_3+1		N_2
5	M_3+1		N_2
6	M_5+1		1
7	M_5+1		1
8	M_1+1		N_1
9	M_2+1		N_1
10	M_4		N_3+1
11	M_3		N_2+1
12	M_3		N_2+1
13	M_1		N_1+1
14	M_2		N_1+1

```
MAPBGN(I) = I+1 - (I+6)/8 - I/7 - 1
MAP12(I)  = (I+3)/5 - (I+3)/9 + 1
MUNSEG(I) = (I+2+(I/10))/2
MAPOFF(I) = I+2*((I)/2)
```

(I)	MAPBGN((I))	MAP12((I))
(1)	1	1
(2)	2	2
(3)	4	2
(4)	6	2
(5)	8	2
(6)	10	1
(7)	11	2
(8)	13	2

I	MUNSEG(I)	MAPOFF(I-9)
1	(1)	
2	(2)	
3	(2)	
4	(3)	
5	(3)	
6	(4)	
7	(4)	
8	(5)	
9	(5)	
10	(6)	1
11	(7)	4
12	(7)	5
13	(8)	8
14	(8)	9

Figure 26. Useful Mappings

$$+ \sum_{m=1}^{M_5+1} (M_{m1}^6 \Lambda_{m1}^6 + M_{m1}^7 \Lambda_{m1}^7)$$

$$+ \sum_{n=1}^{N_1} \left[\sum_{m=1}^{M_1+1} M_{mn}^8 \Lambda_{mn}^8(r) + \sum_{m=i}^{M_2+1} M_{mn}^9 \Lambda_{mn}^9(r) \right]$$

$$M_x = \sum_{n=1}^{N_3+1} \sum_{m=1}^{M_4} M_{mn}^{10} \Lambda_{mn}^{10}(r)$$

$$+ \sum_{n=1}^{N_2+1} \sum_{m=1}^{M_3} (M_{mn}^{11} \Lambda_{mn}^{11}(r) + M_{mn}^{12} \Lambda_{mn}^{12}(r)) \qquad (3.26b)$$

$$+ \sum_{n=1}^{N_1+1} \left[\sum_{m=1}^{M_1} M_{mn}^{13} \Lambda_{mn}^{13}(r) + \sum_{m=1}^{M_2} M_{mn}^{14} \Lambda_{mn}^{14}(r) \right]$$

where the M_{mn}^i's are the unknowns and

$$\Lambda_{mn}^i = \Lambda_m^i(x) \Pi_{n-1/2}^i(y) \qquad i = 10, 11, \ldots, 14$$
$$\Lambda_{mn}^i = \Pi_{m-1/2}^i(x) \Lambda_n^i(y) \qquad i = 1, 2, \ldots, 9 \qquad (3.27)$$

Note that

$$\Pi_{i-1/2}^j(z) = \begin{cases} 1 & z_{i-1}^j < z < z_i^j \\ 0 & \text{otherwise} \end{cases} \qquad (3.28)$$

and

$$\Lambda_i^j(z) = \begin{cases} 1 - \dfrac{z_i^j - z}{z_i^j - z_{i-1}^j} & z_{i-1}^j < z < z_i^j \\ 1 - \dfrac{z - z_i^j}{z_{i+1}^j - z_i^j} & z_i^j < z < z_{i+1}^j \\ 0 \quad \text{otherwise} & \end{cases} \tag{3.29}$$

The x_i and $x_{i-1/2}$ terms are shown in Figure 24. The superscript denotes the region number. This current expansion is convenient for determining

$$\nabla \cdot \vec{M}_s = \frac{\partial M_x}{\partial_x} + \frac{\partial M_y}{\partial_y} \tag{3.30}$$

Eq. (3.26) can also be written in the form

$$M_y = \sum_{i=1}^{9} \sum_{n=1}^{N_i} \sum_{m=1}^{M_i+1} M_{mn}^i \, \Pi_{m-1/2}^i (x) \, \Lambda_n^i (y) \tag{3.31a}$$

$$M_x = \sum_{i=10}^{14} \sum_{n=1}^{N_i+1} \sum_{m=1}^{M_i} M_{mn}^i \, \Lambda_m^i (x) \, \Pi_{n-1/2}^i (y) \tag{3.31b}$$

The $\partial M_y/\partial_y$ and $\partial M_x/\partial_x$, therefore, contain derivatives of the triangle function. It can be shown that

$$\frac{\partial \Lambda_m^i(x)}{dx} = \frac{1}{\Delta x^i} \left\{ \Pi_{m-1/2}^i (x) - \Pi_{m+1/2}^i (x) \right\} \tag{3.32a}$$

$$\frac{\partial \Lambda_n^i(y)}{dy} = \frac{1}{\Delta y^i} \left\{ \Pi_{n-1/2}^i (y) - \Pi_{n+1/2}^i (y) \right\} \tag{3.32b}$$

Therefore,

$$\nabla \cdot \vec{M}_s = \sum_{i=1}^{9} \sum_{n=1}^{N_i} \sum_{m=1}^{M_i+1} M^i_{mn} \Pi^i_{m-1/2}(x) \frac{1}{\Delta y^i} \left\{ \Pi^i_{n-1/2}(y) - \Pi^i_{n+1/2}(y) \right\} \tag{3.33}$$

$$+ \sum_{i=10}^{14} \sum_{n=1}^{N_i+1} \sum_{m=1}^{M_i} M^i_{mn} \frac{1}{\Delta x^i} \left\{ \Pi^i_{m-1/2}(x) - \Pi^i_{m+1/2}(x) \right\} \Pi^i_{n-1/2}(y)$$

The testing paths were chosen to be

$$\begin{aligned} t^i(x,y) &= \delta(x - x^i_{m-1/2})\, \Pi^i_n(y) \qquad i = 1, 2, \ldots, 9 \\ t^i(x,y) &= \Pi^i_m(x)\, \delta(y - y^i_{n-1/2}) \qquad i = 10, 11, \ldots, 14 \end{aligned} \tag{3.34}$$

Assuming $\vec{F}(x,y,0) = \vec{F}(x,y)$ and $\psi(x,y,0) = \psi(x,y)$, the tested integral equation can be shown to be

$$\begin{aligned} &\omega\epsilon_1 F_{s1y}(x^i_{m-1/2}, y^i_n)\Delta y^i + \omega\epsilon_0 F_{s0y}(x^i_{m-1/2}, y^i_n)\Delta y^i + \\ &+ \frac{1}{\omega\mu_1} \{\psi_{s1}(x^i_{m-1/2}, y^i_{n+1/2}) - \psi_{s1}(x^i_{m-1/2}, y^i_{n-1/2})\} + \\ &+ \frac{1}{\omega\mu_0} \{\psi_{s0}(x^i_{m-1/2}, y^i_{n+1/2}) - \psi_{s0}(x^i_{m-1/2}, y^i_{n-1/2})\} = \\ &= \frac{1}{2j} H^{sci}_{0y}(x^i_{m-1/2}, y^i_n)\Delta y^i \end{aligned} \tag{3.35a}$$

$$\begin{aligned} i &= 1, 2, \ldots, 9 \\ m &= 1, 2, \ldots, M_i+1 \\ n &= 1, 2, \ldots, N_i \end{aligned}$$

$$\omega\epsilon_1 F_{s1x}(x^i_m, y^i_{n-1/2})\Delta x^i + \omega\epsilon_0 F_{s0x}(x^i_m, y^i_{n-1/2})\Delta x^i +$$

$$+ \frac{1}{\omega\mu_1}(\psi_{s1}(x^i_{m+1/2}, y^i_{n-1/2}) - \psi_{s1}(x^i_{m-1/2}, y^i_{n-1/2})) +$$

$$+ \frac{1}{\omega\mu_0}(\psi_{s0}(x^i_{m+1/2}, y^i_{n-1/2}) - \psi_{s0}(x^i_{m-1/2}, y^i_{n-1/2})) =$$

$$= \frac{1}{2j} H^{sci}_{0x}(x^i_m, y^i_{n-1/2})\Delta x^i \qquad (3.35b)$$

$$i = 10, 11, 12, 13, 14$$
$$m = 1, 2, \ldots, M_i$$
$$n = 1, 2, \ldots, N_i+1$$

After applying the expansion functions to the tested equation, and approximating the triangle functions, where possible, by pulses, one obtains the result

$$\sum_{j=1}^{9}\sum_{q=1}^{N_j}\sum_{p=1}^{M_j+1} A^{ij}_{mn,pq} M^j_{pq} + \sum_{j=10}^{14}\sum_{q=1}^{N_j+1}\sum_{p=1}^{M_j} B^{ij}_{mn,pq} M^j_{pq} = H^i_{ymn} \qquad (3.36a)$$

for $i = 1, 2, \ldots, 9$; $m = 1, 2, \ldots, M_i+1$; $n = 1, 2, \ldots, N_i$ and

$$\sum_{j=1}^{9}\sum_{q=1}^{N_j}\sum_{p=1}^{M_j+1} C^{ij}_{mn,pq} M^j_{pq} + \sum_{j=10}^{14}\sum_{q=1}^{N_j+1}\sum_{p=1}^{M_j} D^{ij}_{mn,pq} M^j_{pq} = H^i_{xmn} \qquad (3.36b)$$

for $i = 10, 11, 12, 13, 14$; $m = 1, 2, \ldots, M_i$; $n = 1, 2, \ldots, N_i+1$.

The following expressions hold for A, B, C, and D:

For $i = 1, 2, \ldots, 9$ and $j = 1, 2, \ldots, 9$, one obtains

$$A^{ij}_{mn,pq} = \omega\epsilon_1(y^i_{n-1/2} - y^i_{n-1/2})\,\Phi^j_{p,q+1/2}(k_1|x^i_{m-1/2}, y^i_n)$$

$$+ \omega\epsilon_0(y^i_{n+1/2} - y^i_{n-1/2})\,\Phi^j_{p,q+1/2}(k_0|x^i_{m-1/2}, y^i_n)$$

$$+ \frac{1}{\omega\mu_1}\left\{\frac{\Phi^i_{pq}(k_1|x^i_{m-1/2}, y^i_{n+1/2}) - \Phi^j_{pq}(k_1|x^i_{m-1/2}, y^i_{n-1/2})}{y^j_q - y^j_{q-1}}\right.$$

$$
- \frac{\Phi^j_{p,q+1}(k_1|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{p,q+1}(k_1|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_{q+1} - y^j_q} \Bigg\}
$$

$$
+ \frac{1}{\omega\mu_0} \Bigg\{ \frac{\Phi^j_{pq}(k_0|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{pq}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_q - y^j_{q-1}}
$$

$$
- \frac{\Phi^j_{p,q+1}(k_0|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{p,q+1}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_{q+1} - y^j_q} \Bigg\} \tag{3.37a}
$$

Similarly, for i = 1, 2, ..., 9 and j = 10, 11, 12, 13, 14, one obtains

$$
B^{ij}_{mn,pq} = \frac{1}{\omega\mu_1} \Bigg\{ \frac{\Phi^j_{pq}(k_1|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{pq}(k_1|x^i_{m-1/2},\, y^i_{n-1/2})}{x^j_p - x^j_{p-1}}
$$

$$
- \frac{\Phi^j_{p+1,q}(k_1|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{p+1,q}(k_1|x^i_{m-1/2},\, y^i_{n-1/2})}{x^j_{p+1} - x^j_p} \Bigg\}
$$

$$
+ \frac{1}{\omega\mu_0} \Bigg\{ \frac{\Phi^j_{pq}(k_0|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{pq}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{x^j_p - x^j_{p-1}} \tag{3.37b}
$$

$$
- \frac{\Phi^j_{p+1,q}(k_0|x^i_{m-1/2},\, y^i_{n+1/2}) - \Phi^j_{p+1,q}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{x^j_{p+1} - x^j_p} \Bigg\}
$$

For i = 10, 11, 12, 13, 14 and j = 1, 2, ..., 9, one obtains

$$
C^{ij}_{mn,pq} = \frac{1}{\omega\mu_1} \Bigg\{ \frac{\Phi^j_{pq}(k_1|x^i_{m+1/2},\, y^i_{n-1/2}) - \Phi^j_{pq}(k_1|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_q - y^j_{q-1}}
$$

$$
- \frac{\Phi^j_{p,q+1}(k_1|x^i_{m+1/2},\, y^i_{n-1/2}) - \Phi^j_{p,q+1}(k_1|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_{q+1} - y^j_q} \Bigg\}
$$

$$
+ \frac{1}{\omega\mu_0} \Bigg\{ \frac{\Phi^j_{pq}(k_0|x^i_{m+1/2},\, y^i_{n-1/2}) - \Phi^j_{pq}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_q - y^j_{q-1}} \tag{3.37c}
$$

$$
- \frac{\Phi^j_{p,q+1}(k_0|x^i_{m+1/2},\, y^i_{n-1/2}) - \Phi^j_{p,q+1}(k_0|x^i_{m-1/2},\, y^i_{n-1/2})}{y^j_{q+1} - y^j_q} \Bigg\}
$$

Finally, for i = 10, 11, 12, 13, 14 and j = 10, 11, 12, 13, 14, one obtains

$$
\begin{aligned}
D^{ij}_{mn,pq} &= \omega\epsilon_1 (x^i_{m+1/2} - x^i_{m-1/2})\, \Phi^j_{p+1/2,q}(k_1 | x^i_m, y^i_{n-1/2}) \\
&+ \omega\epsilon_0 (x^i_{m+1/2} - x^i_{m-1/2})\, \Phi^j_{p+1/2,q}(k_0 | x^i_m, y^i_{n-1/2}) \\
&+ \frac{1}{\omega\mu_1} \left\{ \frac{\Phi^j_{pq}(k_1 | x^i_{m+1/2}, y^i_{n-1/2}) - \Phi^j_{pq}(k_1 | x^i_{m-1/2}, y^i_{n-1/2})}{x^j_p - x^j_{p-1}} \right. \\
&\left. - \frac{\Phi^j_{p+1,q}(k_1 | x^i_{m+1/2}, y^i_{n-1/2}) - \Phi^j_{p+1,q}(k_1 | x^i_{m-1/2}, y^i_{n-1/2})}{x^j_{p+1} - x^j_p} \right\} \\
&+ \frac{1}{\omega\mu_0} \left\{ \frac{\Phi^j_{pq}(k_0 | x^i_{m+1/2}, y^i_{n-1/2}) - \Phi^j_{pq}(k_0 | x^i_{m-1/2}, y^i_{n-1/2})}{x^j_p - x^j_{p-1}} \right. \\
&\left. - \frac{\Phi^j_{p+1,q}(k_0 | x^i_{m+1/2}, y^i_{n-1/2}) - \Phi^j_{p+1,q}(k_0 | x^i_{m-1/2}, y^i_{n-1/2})}{x^j_{p+1} - x^j_p} \right\}
\end{aligned}
\tag{3.37d}
$$

The potential function common to the A, B, C, and D terms is

$$
\Phi^j_{pq}(k_a | x^i_{m-1/2}, y^i_{n-1/2}) = \int_{x^i_{m-1/2}-x^j_{p-1}}^{x^i_{m-1/2}-x^j_p} \int_{y^i_{n-1/2}-y^j_{q-1}}^{y^i_{n-1/2}-y^j_q} G(k_a; x, y)\, dx dy \tag{3.38}
$$

where a = 0 or 1 for medium 0 or medium 1, respectively, and

$$
G(k_a; x, y) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} e^{-jk_a u_0 r D_x} e^{-jk_a v_0 s D_y} \frac{e^{-jk_a \sqrt{(rD_x - x)^2 + (sD_y - y)^2}}}{4\pi \sqrt{(rD_x - x)^2 + (sD_y - y)^2}} \tag{3.39}
$$

D. Series Acceleration

Since the doubly infinite series kernel of Eq. (3.39) is slowly converging, a series acceleration technique is used. The series acceleration uses a subtraction method [28]. It involves the two dimensional Poisson's transformation of the form

$$\sum_m \sum_n g_0(m,n) = \sum_p \sum_q G_0(2\pi p, 2\pi q) \tag{3.40}$$

where $g_0(\alpha,\beta)$ and $G_0(\omega_A,\omega_B)$ are Fourier transforms of each other. Consider

$$g_0(\alpha,\beta) = e^{-jk_{x\alpha}(x-x_0')} \; e^{-jk_{y\beta}(y-y_0')} \; \frac{e^{-K_{\alpha\beta}|z|}}{K_{\alpha\beta}} \tag{3.41}$$

where

$$\begin{aligned} k_{x\alpha} &= k(u_0 + \alpha \frac{\lambda}{D_x}) \\ k_{y\beta} &= k(v_0 + \beta \frac{\lambda}{D_y}) \\ K_{\alpha\beta}^2 &= k_{x\alpha}^2 + k_{y\beta}^2 + u^2 \end{aligned} \tag{3.42}$$

The Fourier transform of $g(\alpha,\beta)$ is

$$G_0(\omega_A,\omega_B) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_0(\alpha,\beta) \; e^{-j\alpha\omega_A} \; e^{-j\beta\omega_B} \; d\alpha d\beta = \frac{D_x D_y}{2\pi} \; e^{j\frac{D_x}{\lambda}u_0\omega_A} \; e^{j\frac{D_y}{\lambda}v_0\omega_B} \; \frac{e^{-uS}}{S} \tag{3.43}$$

where

$$S^2 = (x-x_0' + \frac{D_x}{2\pi}\omega_A)^2 + (y-y_0' + \frac{D_y}{2\pi}\omega_B)^2 + z^2 \tag{3.44}$$

Evaluating Eq. (3.41) at $\alpha=m$, $\beta=n$, and Eq. (3.43) at $\omega_A=2\pi p$ and

$\omega_B = 2\pi q$, one obtains

$$g_0(m,n) = e^{-jk_{xm}(x-x_0')} \; e^{-jk_{yn}(y-y_0')} \; \frac{e^{-K_{mn}|z|}}{\Gamma_{mn}} \tag{3.45}$$

where

$$\begin{aligned} k_{xm} &= k(u_0 + m\frac{\lambda}{D_x}) \\ k_{yn} &= k(v_0 + n\frac{\lambda}{D_y}) \\ K_{mn}^2 &= k_{xm}^2 + k_{yn}^2 + u^2 \end{aligned} \tag{3.46}$$

and

$$G_0(2\pi p, 2\pi q) = \frac{D_x D_y}{2\pi} \; e^{j\frac{D_x}{\lambda}u_0 2\pi p} \; e^{j\frac{D_y}{\lambda}v_0 2\pi q} \; \frac{e^{-uS}}{S} \tag{3.47}$$

where

$$S^2 = (x-x_0' + pD_x)^2 + (y-y_0' + qD_y)^2 + z^2 \tag{3.48}$$

Defining

$$\begin{aligned} \vec{r} &= x\hat{x} + y\hat{y} + z\hat{z} \\ \vec{r}' &= (x_0' - pD_x)\hat{x} + (y_0' - qD_y)\hat{y}, \end{aligned} \tag{3.49}$$

one can see that

$$S = |\vec{r} - \vec{r}'| = R_{-p,-q} \tag{3.50}$$

If we let $u = jk$ and $K_{mn} = j\Gamma_{mn}$, then

$$\Gamma_{mn}^2 = k^2 - k_{xm}^2 - k_{yn}^2 \tag{3.51}$$

Combining Eq. (3.40) with Eq. (3.45) through Eq. (3.51), one obtains

$$\sum_m \sum_n e^{-jk_{xm}(x-x_0')} e^{-jk_{yn}(y-y_0')} \frac{e^{-j\Gamma_{mn}|z|}}{j\Gamma_{mn}} =$$

$$= 2D_x D_y \sum_p \sum_q e^{jku_0 pD_x} e^{jkv_0 qD_y} \frac{e^{-jkR_{-p,-q}}}{4\pi R_{-p,-q}} \tag{3.52}$$

Changing variables in the last sum to m = -p, n = -q and then resubstituting p for m and q for n, one obtains the transformation

$$G(k;x-x_0',y-y_0',z) = \sum_p \sum_q e^{-jku_0 pD_x} e^{-jkv_0 qD_y} \frac{e^{-jkR_{pq}}}{4\pi R_{pq}}$$

$$= \frac{1}{2jD_x D_y} \sum_m \sum_n e^{-jk_{xm}(x-x_0')} e^{-jk_{yn}(y-y_0')} \frac{e^{-j\Gamma_{mn}|z|}}{\Gamma_{mn}} \tag{3.53}$$

The series acceleration technique is based on combining Kummer's transformation and the Poisson summation formula given by Eq. (3.53). The technique [28] requires that a smooth function be defined that is asymptotically equal to the original.

Let

$$F(2\pi p, 2\pi q, z) = e^{-jku_0 pD_x} e^{-jkv_0 qD_y} \frac{e^{-jkR_{pq}}}{4\pi R_{pq}}$$

where (3.54)

$$R_{pq}^2 = (x_0' + pD_x - x)^2 + (y_0' + qD_y - y)^2 + z^2$$

Then from Eq. (3.45) and Eq. (3.53), the Fourier transform is

$$f(m,n,z) = \frac{1}{2jD_x D_y} e^{-jk_{xm}(x-x_0')} e^{-jk_{yn}(y-y_0')} \frac{e^{-j\Gamma_{mn}|z|}}{\Gamma_{mn}} \tag{3.55}$$

Now replace z by $\sqrt{z^2+c^2}$ to obtain a new function, where c is an appropriately chosen real number. Adding and subtracting the new functions, one obtains

$$\sum_p \sum_q F(2\pi p, 2\pi q, z) = \sum_p \sum_q \left\{ F(2\pi p, 2\pi q, z) - F(2\pi p, 2\pi q, \sqrt{z^2+c^2}) \right\} + \sum_m \sum_n f(m, n, \sqrt{z^2+c^2}) \tag{3.56}$$

The terms with $\sqrt{z^2+c^2}$ are obtained from Eq. (3.53). Since the integral equation is enforced at z = 0, one obtains

$$\sum_m \sum_n F(2\pi m, 2\pi n, 0) = \sum_m \sum_n D(2\pi m, 2\pi n) + \sum_m \sum_n f(m, n, c) \tag{3.57}$$

where $D(2\pi m, 2\pi n) = F(2\pi m, 2\pi n, 0) - F(2\pi m, 2\pi n, c)$. Since F is given by Eq. (3.54), one finds

$$D(2\pi m, 2\pi n) = \frac{e^{-jku_0 pD_x} e^{-jkv_0 qD_y}}{4\pi} \left\{ \frac{e^{-jk\sqrt{(x-x_0' + mD_x)^2 + (y-y_0' + nD_y)^2}}}{\sqrt{(x-x_0' + mD_x)^2 + (y-y_0' + nD_y)^2}} - \frac{e^{-jk\sqrt{(x-x_0' + mD_x)^2 + (y-y_0' + nD_y)^2 + c^2}}}{\sqrt{(x-x_0' + mD_x)^2 + (y-y_0' + nD_y)^2 + c^2}} \right\} \tag{3.58}$$

The expression for f(m,n,c) is

$$f(m,n,c) = \frac{1}{2jD_x D_y} e^{-jk_{xm}(x-x_0')} e^{-jk_{yn}(y-y_0')} \frac{e^{-j\Gamma_{mn} c}}{\Gamma_{mn}} \tag{3.59}$$

$$\text{where } \Gamma_{mn} = \begin{cases} \sqrt{k^2 - k_t^2} & k > k_t \\ -j\sqrt{k_t^2 - k^2} & k < k_t \end{cases} \tag{3.60}$$

$$k_t^2 = k_{xm}^2 + k_{yn}^2$$

A problem occurs when $x = x_0'$, $y = y_0'$, and $m = n = 0$ in $D(2\pi m, 2\pi n)$. For this reason, $m = n = 0$ is treated separately in the sum.

Separating out the $r = s = 0$ term in Eq. (3.39), one obtains

$$G(k_a;x,y) = \frac{e^{-jk_a\sqrt{x^2+y^2}}}{4\pi\sqrt{x^2+y^2}} + \sum_{\substack{r \\ r=s\neq 0}} \sum_{s} F(2\pi r, 2\pi s, 0) \tag{3.61}$$

Since a problem exists in the first term when $x = y = 0$, it can be split into a part to be integrated analytically and another part to be integrated numerically. Doing this by adding and subtracting identical terms, one obtains

$$G(k_a;x,y) = \frac{\cos k_a\sqrt{(x^2+y^2)} - (1 - k_a^2(\sqrt{(x^2+y^2)})^2/2)}{4\pi\sqrt{(x^2+y^2)}}$$

$$-j\frac{\sin k_a\sqrt{(x^2+y^2)}}{4\pi\sqrt{(x^2+y^2)}} + \sum_{\substack{r \\ r=s\neq 0}} \sum_{s} F(2\pi r, 2\pi s, 0) \tag{3.62}$$

$$+ \frac{(1 - k_a^2(\sqrt{(x^2+y^2)})^2/2)}{4\pi\sqrt{(x^2+y^2)}}$$

The last term was integrated analytically. The other terms were integrated numerically using double integration with second order Gaussian Quadrature. The doubly infinite sum is accelerated by Eq. (3.57). Therefore,

$$\sum_{\substack{r\\ r=s\neq 0}}\sum_{s} F(2\pi r, 2\pi s, 0) = \sum_{\substack{m\\ m=n\neq 0}}\sum_{n} D(2\pi m, 2\pi n) - F(0,0,c) + \sum_{m}\sum_{n} f(m,n,c) \quad (3.63)$$

where $F(0,0,c) = \dfrac{e^{-jk\sqrt{(x-x_0')^2 + (y-y_0')^2 + c^2}}}{4\pi\sqrt{(x-x_0')^2 + (y-y_0')^2 + c^2}}$

The analytic integration becomes

$$\int_{x_a}^{x_b}\int_{y_a}^{y_b} \frac{1 - \frac{(k_a\sqrt{(x^2+y^2)})^2}{2}}{4\pi\sqrt{(x^2+y^2)}}\, dy\, dx =$$

$$= \frac{1}{4\pi}\left\{ x_b\left(1 - \frac{k_a^2 x_b^2}{12}\right)\left[\ln(y_b + \sqrt{(x_b^2+y_b^2)}) - \ln(y_a + \sqrt{(x_b^2+y_a^2)})\right]\right.$$

$$+ x_a\left(1 - \frac{k_a^2 x_a^2}{12}\right)\left[\ln(y_a + \sqrt{(x_a^2+y_a^2)}) - \ln(y_b + \sqrt{(x_a^2+y_b^2)})\right]$$

$$\left. + y_b\left(1 - \frac{k_a^2 y_b^2}{12}\right)\left[\ln(x_b + \sqrt{(x_b^2+y_b^2)}) - \ln(x_a + \sqrt{(x_a^2+y_b^2)})\right]\right. \quad (3.64)$$

$$+ y_a \left(1 - \frac{k_a^2 y_a^2}{12} \right) \left[\ell n(x_a + \sqrt{(x_a^2+y_a^2)}) - \ell n(x_b + \sqrt{(x_b^2+y_a^2)}) \right]$$

$$- \frac{k_a^2 x_b y_b}{6} \sqrt{(x_b^2+y_b^2)} + \frac{k_a^2 x_a y_b}{6} \sqrt{(x_a^2+y_b^2)}$$

$$+ \frac{k_a^2 x_b y_a}{6} \sqrt{(x_b^2+y_a^2)} - \frac{k_a^2 x_a y_a}{6} \sqrt{(x_a^2+y_a^2)}$$

The expressions for Φ were evaluated for all i, j, p, q, m, and n and stored in two matrices, one for k_0 and the other for k_1. The matrix elements A, B, C, and D were obtained from them by proper manipulation of the matrices. This is estimated to cut the matrix element evaluation time by 6. The storage requirement more than doubled, however, to acheive the time savings.

The right-hand side is taken to be zero everywhere except in the feed region where it is taken to be 1/2j. Thus, $H_x^{sci}(x_m^i, y_{n-1/2}^i)\Delta x^i = 1$ over this region.

RESULTS

The program was run twice so far. When matrix elements were evaluated to six digit accuracy, it ran about 30 hours. When they were evaluated to about three digit accuracy, it ran about an hour and forty minutes. A more complete discussion of the results will be given at the URSI meeting in Syracuse in June 1988.

REFERENCES

1. D. F. Hanson, "Fields of a Slot Antenna on a Half-Space Fed by Coplanar Waveguide Using the Method of Moments," 1986 USAF-UES SUMMER FACULTY RESEARCH PROGRAM FINAL REPORT, Contract No. F49620-85-C-0013, August 18, 1986.

2. D. F. Hanson, "A Moment Method Solution for a Slot Antenna Fed by Coplanar Waveguide," 1987 Spring IEEE AP-S/URSI International Symposium Digest, Vol I, Virginia Tech., Blacksburg, VA, June 15-19, 1987, pp. 107-110.

3. D. F. Hanson, "A Study of Coplanar Waveguide and its Application to Phased Arrays of Integrated Circuit Antennas," 1985 USAF-UES SUMMER FACULTY RESEARCH PROGRAM FINAL REPORT, Contract No. F49620-85-C-0013, August 12, 1985.

4. D. F. Hanson, Principal Investigator, Universal Energy Systems/AFOSR Contract No. F49620-85-C-0013/SB5851-0360, "An Infinite Phased Array of Slots Fed by Coplanar Waveguide Over a Dielectric Half-Space," 1987.

5. C. P. Wen, "Coplanar Waveguide: A Surface Strip Transmission Line Suitable for Nonreciprocal Gyromagnetic Device Applications," IEEE Trans. Microwave Theory Techn., Vol. MTT-17, No. 12, Dec. 1969, pp. 1087-1090.

6. K. C. Gupta, R. Garg, and I. J. Bahl, Microstrip Lines and Slotlines, Dedham, MA: Artech House, Inc., 1979.

7. R. Soares, J. Graffeuil and J. Obregon, Applications of GaAs MESFETs, Dedham, MA: Artech House, 1983.

8. R. A. Pucel, "Design Considerations for Monolithic Microwave Circuits," IEEE Trans. Microwave Theory Techn., Vol. MTT-29, June 1981, pp. 513-534.

9. K. J. Button, Infrared and Millimeter Waves, Vol. 10, Millimeter Components and Techniques, Part II. New York: Academic Press, 1983. Chapter 1. "Integrated Circuit Antennas" by D. B. Rutledge, D. P. Neikirk, and D. P. Kasilingam, pp. 1-90.

10. I. J. Bahl and P. Bhartia, Microstrip Antennas, Dedham, MA: Artech House, 1980.

11. IEEE Trans. Antennas Prop., Vol. AP-29, No. 1, January 1981.

12. D. T. McGrath, D. A. Mullinix, and K. D. Huck, "Fortran Subroutines for Design of Printed Circuit Antennas," RADC-TR-86, May 1986.

13. G. Ghione and C. U. Naldi, "Coplanar Waveguides for MMIC Applications: Effect of Upper Shielding, Conductor Backing, Finite-Extent Ground Planes, and Line-to-Line Coupling," IEEE Trans. Microwave Theory Techn., Vol. MTT-35, No. 3, March 1987, pp. 260-267.

14. K. C. Gupta, R. Garg, and R. Chadha, Computer-Aided Design of Microwave Circuits, Dedham, MA: Artech House, Inc., 1981.

15. D. A. Rowe and B Y. Lao, "Numerical Analysis of Shielded Coplanar Waveguides," IEEE Trans. Microwave Theory Techn., Vol. MTT-31, No. 11, Nov. 1983, pp. 911-915.

16. G. Ghione and C. Naldi, "Parameters of Coplanar Waveguides with Lower Ground Plane," Electronics Letters, Vol. 19, No. 18, pp. 734-735, Sept. 1, 1983.

17. M. Kominami, D. M. Pozar, and D. H. Schaubert, "Dipole and Slot Elements and Arrays on Semi-Infinite Substrates," IEEE Trans. Antennas Prop., Vol. AP-33, No. 6, pp. 600-607, June 1985.

18. J. R. Souza and E. C. Talboys, "S-Parameter Characterization of Coaxial to Microstrip Transition," IEE Proc., Vol. 129, Pt. H, No. 1, pp. 37-40, Feb. 1982.

19. D. Kajfez, Notes on Microwave Circuits, Vol. II, Kajfez Consulting Co., Oxford, MS 1986.

20. A. Nesic, "Printed Slot Array Excited by a Coplanar Waveguide," Proceedings of the 12th European Microwave Conference, pp. 478-482, Helsinki, Finland, September 13-16, 1982.

21. R. F. Harrington, Time-Harmonic Electromagnetic Fields, McGraw-Hill Book Co., New York, 1961.

22. R. J. Mailloux, "On the Use of Metallized Cavities in Printed Slot Arrays with Dielectric Substrates," IEEE Trans. Antennas Prop., Vol. AP-35, No. 5, pp. 477-487, May 1987.

23. F. M. Arscott, Periodic Differential Equations: An Introduction to Mathiew, Lame, and Allied Functions, Macmillan Co., New York, 1964.

24. W. Magnus and F. Oberhettinger, Formulas and Theorems for the Special Functions of Mathematical Physics, Chelsea, New York, 1949.

25. N. Amitay, V. Galindo, and C. P. Wu, Theory and Analysis of Phased Array Antennas, Wiley-Interscience, New York, 1972.

26. E. Argence and T. Kahan, Theory of Waveguides and Cavity Resonators, Hart Publishing Co., New York, 1968.

27. C. M. Butler, D. R. Wilton, and A. W. Glisson, Notes from a Short Course on "Fundamentals of Numerical Solution Methods in Electromagnetics," University of Mississippi, Oxford, MS 1982.

28. R. Lampe, P. Klock, and P. Mayes, "Integral Transforms Useful for the Accelerated Summation of Periodic, Free-Space Green's Functions," IEEE Microwave Theory and Techn., Vol. MTT-33, No. 8, pp. 734-736, August 1985.

A New Measure of Maintainability/Reliability and Its Estimation

J. Marcus Jobe

Miami University, Oxford, Ohio

Key Words - Force of mortality, Maintenance support burden, Maximum likelihood estimation

Reader Aids -

Purpose: Report of a new measure of maintainability/reliability and its estimation

Special math needed for explanation: Statistics

Special math needed to use results: Same

Results useful to: Reliability engineers and theorists

Abstract - A maintainability/reliability measure discussed in this paper is referred to as MTUT. It corresponds to the average time to restore an equipment and maintenance system to its original working status expressed as a proportion of the mean time to failure for any given equipment. This measure integrates maintenance and repair time expenditures of all types from three levels of maintenance. Other measures discussed in the literature such as availability (A), mean time to repair (MTTR), and average queue length (ρ) are compared to MTUT. Further, a testing program for the demonstration phase of equipment development is presented. Estimation and discrimination procedures are derived for MTUT using data from the outlined testing program. Large sample theory is used to construct both interval estimates and discrimination procedures for the MTUT parameter using data acquired from the assessment phase of equipment development.

1. INTRODUCTION

Equipment which supports the operations of military interests must have the capability to consistently perform its intended tasks under various extreme conditions. Upon breakdown, the design of an equipment partially determines the time necessary to restore the item to functional status. The ability to identify designs producing a minimum average maintenance time for a given operating time would be beneficial for military concerns. We propose methods of accomplishing this goal in what follows.

Common measures of maintainability/reliability consider only the first level of maintenance and the corresponding force of mortality or rate of occurrence of failures. The most common of these measures are availability (A), mean time to repair (MTTR), and mean queue length (ρ). The repair times required to return the maintenance system to its original state are not included in these measures. If the additional repair times were incorporated into these measures, their respective interpretations would be different than currently perceived.

In the military, the maintenance on a failed equipment affects several levels of operation. The system or equipment may be restored to an operating status but the maintenance required for the correction of a failed module is not necessarily completed. The ripple effect created over the various echelons of a maintenance system can be thought of as the reflection of a maintenance support burden. Specifically, the maintenance support burden will be defined as the maintenance time required to return the equipment and maintenance system to its original state for a given period of equipment operating time. A new measure of maintainability/reliability which takes into account the repair times at the preassigned maintenance stages will be the focus of this paper. We will refer to this measure as the average time

required to restore an equipment and maintenance system to its original operating state expressed as a proportion of the average lifetime for a given equipment type.

Information from three steps of a maintenance system (LRU, SRU, and circuit) will be considered. A brief comparison of A, MTTR, and ρ to MTUT is discussed in section 3. Section 4 outlines a testing program useful during the demonstration phase of an equipment. The respective estimation and discrimination procedure for MTUT (when the force of mortality is known for each component making up an equipment) is given for this testing program. Estimation and testing procedures for MTUT with unknown force of mortality are in section 5. Mathematical derivations referred to in sections 4 and 5 are included in appendices A, B, and C. Future work is mentioned in section 6. A summary section concludes the paper.

2. ASSUMPTIONS AND NOTATION

Assumptions

1. All equipment considered in this paper are made up of LRUs in series. SRUs are in series within each LRU and circuits are in series within each SRU. Hence, any equipment failure can be traced to a single circuit malfunction.

2. The component failure times have s-independent exponential distributions with possibly different means for each type of component.

3. The respective repair times for the component types have s-independent exponential distributions with possibly different means.

4. Failure times are s-independent of repair times regardless of the component under consideration.

Notation and Nomenclature

LRU	line replaceable unit
SRU	shop replaceable unit
λ	force of mortality of time to failure
	i,j,l,: indices for LRU, SRU, and circuit
λ_T	force of mortality of time to failure for a single equipment equal to $\sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \lambda_{ijl}$
γ	constant rate of occurrence of failures for a continuous stream of equipment breakdowns occurring in a Poisson fashion
MTTF	mean time to failure for any single equipment, equal to $1/\lambda_T$
Mct_i	average time to locate, remove, and replace a failed LRU
Mct_i^*	average time to disassemble the i^{th} LRU, identify, remove, and replace the failed SRU in the i^{th} LRU and reassemble the i^{th} LRU
Mct_{ij}	average time to disassemble the failed SRU, identify, remove, and replace the faulty circuit in the j^{th} SRU and reassemble the j^{th} SRU in the ith LRU
Mct_{ijl}	average time to disassemble the faulty circuit and correct or fix the l^{th} circuit in the jth SRU from the ith LRU
k	the number of different LRUs in the system or piece of equipment
k_i	the number of different SRUs in the i^{th} LRU (i = 1, 2, ... k)
k_{ij}	the number of circuits in the j^{th} SRU in the i^{th} LRU (j = 1, ... k_i)
n	number of observed equipment failures
	i,j,l,: indices for LRU, SRU, and circuit

$$n = \sum_{i=1}^{k} n_i, \quad n_i = \sum_{j=1}^{k_i} n_{ij}, \quad n_{ij} = \sum_{l=1}^{k_{ij}} n_{ijl}$$

y_{ijlm} the time to locate, remove, and replace the i^{th} failed LRU for the m^{th} occurrence of the l^{th} circuit failure in the j^{th} SRU

x_{ijlm} the repair time of the i^{th} LRU for the m^{th} occurrence of the l^{th} circuit failure in the j^{th} SRU

x^{*}_{ijlm} the repair time of the j^{th} SRU in the i^{th} LRU for the m^{th} occurrence of the l^{th} circuit failure

x^{**}_{ijlm} the m^{th} repair time of the l^{th} circuit in the j^{th} SRU in the i^{th} LRU

t_{ijlm} the m^{th} time to failure of the l^{th} circuit in the j^{th} SRU in the i^{th} LRU

$z_{\alpha/2}$ value of z such that gauf(z) = $1-\alpha/2$

3. MTTR, MTUT, ρ, and A

Considering only the maintenance time required to restore an equipment to working status, identities for MTTR, MTUT, A, and ρ are given below

$$MTTR \equiv \sum_{i=1}^{k} (\lambda_i/\lambda_T)\, Mct_i. \tag{1}$$

$$MTUT \equiv \sum_{i=1}^{k} \lambda_i Mct_i = MTTR/MTTF \tag{2}$$

$$\rho \equiv \gamma \cdot MTTR = \text{mean service time/mean inter-arrival time between equipment failures} \tag{3}$$

$$A \equiv 1/(1 + MTTR/MTTF). \tag{4}$$

The average time required to restore an equipment and maintenance system to its original state expressed as a proportion of the average lifetime (MTUT) reflects the mean impact any single equipment failure has on the maintenance

system. The average time to restore a single equipment to an operating status is MTTR, and the steady state probability that a single equipment is operating satisfactorily at any point in time is availability (A). In contrast, according to Parzen [5], the mean queue length (ρ) represents the impact a continuous stream of equipment failures (occurring in Poisson fashion with constant rate of occurrence of failures, γ) has on a maintenance system.

The distinction between a "single" equipment and "continuous stream" of equipments is important. MTTF applies to any one equipment, whereas $1/\gamma$ is the mean interarrival time of a continuous stream of failed equipments (not necessarily equal to MTTF). A detailed discussion of the important difference between the force of mortality of time to failure and the rate of occurrence of failures in a continuous stream can be found in Ascher [1], Ascher and Feingold [2], and Thompson [6]. We will not concern ourselves with a stream of occurrences of failures in this paper; instead, the focus is on a lifetime and maintenance time for a single equipment type. Hence, MTUT and ρ do not represent the same concept (nor are they necessarily equal). If only the maintenance time needed to restore an equipment to functional status is considered and γ is known, ρ can be computed from MTTR. We turn now to examples illustrating shortcomings of MTTR as a measure of maintenance support burden.

The measure MTTR is of importance at the LRU level of maintenance as an indicator of how quickly on the average an equipment or system can be returned to operation given a failure has occurred. The following example is given in Klion [4]. It is given here to help motivate the use of MTUT as a more informative and interpretable measure of maintenance support burden than MTTR. Corrective maintenance for a single maintenance echelon, say LRU, is considered in this example for clarity. At the end of this section, an

example is given to illustrate the computation of MTUT when two stages of maintenance at the LRU echelon are considered.

Example 1

An equipment comprised of four repairable/removable modules in series is to be considered. Figures I, II, and III give values of λ_i and Mct_i for three systems of this type. Figures II and III reflect modifications of the system described by Figure I. We note that the maintainability design characteristics (Mct_i) remain unchanged for the respective modules in each of these systems.

The value of the MTTR measure has deteriorated (increased) for the modified design given by Figure II while the value of MTUT (a measure reflecting maintenance support burden) has improved (decreased). The apparent conflict is further amplified when we realize that the reliability of the equipment described in Figure II has increased over the reliability of the equipment depicted by Figure I.

Turning our attention to the design depicted in Figure III, the forces of mortality for modules (1) and (2) have increased as a result of an enhancement to possibly increase performance. The resulting reliability of the system has decreased, but the MTTR measure is more attractive (smaller) relative to the original equipment. We see that the value of MTUT has increased considerably, however, for the altered design.

We can think of a situation where all forces of mortality for the respective modules in an equipment are reduced by a factor of 1/2. The value of MTTR for the new design does not change but the MTUT value is decreased to half its original value. We also note that the reliability of the altered system is increased.

If the original equipment is changed such that the force of mortality for all modules are doubled, the MTTR value of the new equipment remains unchanged. The MTUT value doubles while the reliability of the new equipment significantly decreases.

$\lambda_1 = .003333$ $Mct_1 = 1/2$	$\lambda_2 = .001667$ $Mct_2 = 2/3$
$\lambda_3 = .001111$ $Mct_3 = 1$	$\lambda_4 = .0008333$ $Mct_4 = 2$

MTTR = .8, MTUT = .00556

Fig. I. Equipment containing 4 modules in series with respective forces of mortality and average repair times.

$\lambda_1 = .001667$ $Mct_1 = 1/2$	$\lambda_2 = .0008333$ $Mct_2 = 2/3$
$\lambda_3 = .001111$ $Mct_3 = 1$	$\lambda_4 = .0008333$ $Mct_4 = 2$

MTTR = .94, MTUT = .0041667

Fig. II. Equipment from Fig. I. with decreased forces of mortality for modules 1 and 2.

$\lambda_1 = .006667$ $Mct_1 = 1/2$	$\lambda_2 = .005$ $Mct_2 = 2/3$
$\lambda_3 = .001111$ $Mct_3 = 1$	$\lambda_4 = .0008333$ $Mct_4 = 2$

MTTR = .69, MTUT = .0094

Fig. III. Equipment from Fig. I. with increased forces of mortality for modules 1 and 2.

As pointed out by Klion [4], this example illustrates dilemmas which can occur when using MTTR, or its estimate, to evaluate an equipment's maintainability. We see from these examples that MTUT provides information about the maintainability of an equipment useful for determining the required maintenance support, whereas the MTTR measure can be very misleading in this respect.

The availability measure also has shortcomings as a measure of maintenance support burden. Asher and Feingold [2] list conditions necessary for equation (4) to be equal to the steady state availability. If only the first level of maintenance is considered, conditions (as outlined in the assumptions) are such that equation (4) is the steady state availability. If all levels of maintenance are considered, lifetime and repair time do not constitute an alternating renewal process, thus equation (4) would not be the steady state availability (A). In contrast to the measures discussed (A, MTTR, and ρ), the interpretation of MTUT remains attractive as a measure

consistent with the insight needed to make decisions concerning maintenance support burden when all levels of maintenance are considered.

We conclude this section with an example illustrating the application of MTUT for two stages of maintenance at the LRU level.

Example 2

An equipment comprised of four repairable/removable modules in series is again considered. Figures IV, V, and VI give values of λ_i, Mct_i, and Mct_i^* for three equipments of this type. Figure V reflects a slight modification of IV and VI is a modification of V. The change depicted in V results in an increased equipment reliability and increased MTUT. The equipment portrayed in VI has an increased reliability and decreased MTUT compared to the equipment in V.

$\lambda_1 = .1$ $Mct_1 = 1$ $Mct_1^* = 1$	$\lambda_2 = .2$ $Mct_2 = 2$ $Mct_2^* = 2$
$\lambda_3 = .3$ $Mct_3 = 3$ $Mct_3^* = 3$	$\lambda_4 = .4$ $Mct_4 = 4$ $Mct_4^* = 4$

MTUT = 6.0, $\lambda_T = 1.0$

Fig. IV. Equipment containing 4 modules in series with respective forces of mortality and average repair times for two stages of maintenance at the LRU level.

$\lambda_1 = .1$ $Mct_1 = 1$ $Mct_1^* = 1$	$\lambda_2 = .2$ $Mct_2 = 2$ $Mct_2^* = 2$
$\lambda_3 = .2$ $Mct_3 = 3$ $Mct_3^* = 3$	$\lambda_4 = .49$ $Mct_4 = 4$ $Mct_4^* = 4$

MTUT = 6.12, $\lambda_T = .99$

Fig. V. Equipment from Fig. IV. with a modified force of mortality for modules 3 and 4.

$\lambda_1 = .1$ $Mct_1 = 1$ $\star$ $Mct_1 = 1$	$\lambda_2 = .2$ $Mct_2 = 2$ $\star$ $Mct_2 = 2$
$\lambda_3 = .2$ $Mct_3 = 3$ $\star$ $Mct_3 = 3$	$\lambda_4 = .48$ $Mct_4 = 4$ $\star$ $Mct_4 = 4$

MTUT = 6.04, $\lambda_T = .98$

Fig. VI. Equipment from Fig. V. with a modified force of mortality for module 4.

We turn now to the estimation of MTUT in both the demonstration and assessment stages of equipment development.

4. ESTIMATION AND TESTING FOR MTUT WHEN λ_{ijl} KNOWN

We can write the expression for MTUT when considering the three levels of maintenance (LRU, SRU, and circuit) as

$$MTUT \equiv \sum_{i=1}^{k} \lambda_i Mct_i + \sum_{i=1}^{k} \lambda_i Mct_i^{\star} + \sum_{i=1}^{k} \sum_{j=1}^{k_i} \lambda_{ij} Mct_{ij}$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \lambda_{ijl} Mct_{ijl}. \qquad (5)$$

Because we are assuming constant forces of mortality and independence of the failure times, we see that

$$\lambda_T \equiv \sum_{i=1}^{k} \lambda_i \equiv \sum_{i=1}^{k} \sum_{j=1}^{k_i} \lambda_{ij} \equiv \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \lambda_{ijl}. \qquad (6)$$

Rewriting, (5) becomes

$$MTUT \equiv \lambda_T \cdot \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} (\lambda_{ijl}/\lambda_T)(Mct_i + Mct_i^{*} + Mct_{ij} + Mct_{ijl}). \qquad (7)$$

In order to estimate MTUT, we will assume throughout the following two sections that any particular equipment malfunction can be traced to a single circuit failure. Time constraints usually prohibit the observation of repair times from all modes in the demonstration phase of an equipment. The following discussion outlines the construction of a $(1-\alpha)100\%$ confidence interval using a testing program involving as few as 30 equipment failures, regardless of the number of LRUs, SRUs, and circuits in an equipment. This will result in as few as 120 repair times needed to evaluate the MTUT for the whole equipment.

A single faulty circuit, selected in a probabilistic fashion, is inserted into at least 30 randomly selected equipments of interest. The selected faulty circuit will most likely be different from one equipment to the next. This "failed" circuit corresponds to an inoperative SRU, in turn determining a failed LRU, and thus, a breakdown of the equipment. The values of y_{ijlm}, x_{ijlm}^{*}, x_{ijlm}^{**}, are observed and recorded for each of the 30 "broken" equipments in a "bench repair" setting. An unbiased estimate for MTUT becomes

$$\hat{MTUT} = \lambda_T \cdot \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} (y_{ijlm} + x_{ijlm}^{*} + x_{ijlm}^{**})/30. \qquad (8)$$

We see by the central limit theorem that for a large number of equipment repairs (≥ 30) that $\hat{MTUT} \sim N(MTUT, Var(\hat{MTUT}))$. Therefore, a $(1-\alpha)100\%$ interval estimate of MTUT is,

$$\hat{MTUT} \pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{MTUT})}. \tag{9}$$

An expression for $\hat{Var}(\hat{MTUT})$ is given in Appendix A. It should be noted the identification of the faulty circuit to be inserted in an equipment is important. The 1^{th} circuit in the j^{th} SRU within the i^{th} LRU has a probability of being selected equal to λ_{ijl}/λ_T. Thus, many circuit repairs may not be included in the demonstration, and the ones observed may occur only once. The robustness of the central limit theorem and the nature of the sampling (probability sampling) insures the normality and unbiasedness of equation (8). Decision procedures involving a hypothesized value of MTUT are straightforward for this type of demonstration program using expression (9) (i.e. for the test H_0: $MTUT \geq C_0$ vs. H_A: $MTUT < C_0$, if $\hat{MTUT} + z_\alpha \sqrt{\hat{Var}(\hat{MTUT})} < C_0$, we conclude H_A: $MTUT < C_0$, with a type I error probability being α).

All repair modes usually occur during the assessment stage of a piece of equipment. When the number of repairs grows large for each circuit type ($n_{ijl} \geq 30$), an unbiased, normally distributed estimate of MTUT becomes

$$\hat{MTUT} = \sum_{i=1}^{k} \lambda_i(\bar{y}_i + \bar{x}_i) + \sum_{i=1}^{k}\sum_{j=1}^{k_i} \lambda_{ij}\, \bar{x}^*_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{k_i}\sum_{l=1}^{k_{ij}} \lambda_{ijl}\, \bar{x}^{**}_{ijl}. \tag{10}$$

The computational details for $\bar{y}_i$, $\bar{x}_i$, $\bar{x}^*_{ij}$, and $\bar{x}^{**}_{ijl}$ are given in Appendix A. The resulting $(1-\alpha)100\%$ confidence interval for MTUT at the assessment stage is,

$$\hat{MTUT} \pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{MTUT})}. \tag{11}$$

where $\hat{MTUT}$ comes from (10) and $\hat{Var}(\hat{MTUT})$ is derived in Appendix A. Decision procedures for a hypothesized value of MTUT are the same as that described for the demonstration phase with expression (11) used instead of (9).

5. ESTIMATION AND TESTING FOR MTUT WHEN λ_{ij1} UNKNOWN

Two scenarios to be considered in this section are small sample sizes ($0 < n_{ij1} \leq 30$) and large sample sizes ($n_{ij1} > 30$) for each failure mode of a system. All estimation and testing procedures developed in this section are conditional on n_i, n_{ij}, and n_{ij1} values ($\sum_{i=1}^{k} n_i$, $\sum_{j=1}^{k_i} n_{ij}$, and $\sum_{l=1}^{k_{ij}} n_{ij1}$ are not necessarily equal to n, n_i, and n_{ij}, respectively). The following discussion presents a conservative $(1-\alpha)100\%$ confidence interval and test procedure for MTUT when finite sample sizes for the respective components are observed.

The independence assumptions as well as the assumed distribution of repair and failure times given in section 2 imply that

$$[(Mct_i\lambda_i) \cdot (\bar{t}_i) / \bar{y}i] \sim F_{2n_i,\, 2n_i}.$$

We see that

$$[(F_{2n_i,2n_i,\alpha/2} \cdot \bar{y}_i)/\bar{t}_i,\ (F_{2n_i,2n_i,1-\alpha/2} \cdot \bar{y}_i)/\bar{t}_i] \text{ is}$$

a $(1-\alpha)100\%$ confidence interval for $\lambda_i Mct_i$. Hence, a conservative $(1-\alpha)100\%$ confidence interval for $\sum_{i=1}^{k} \lambda_i Mct_i$ is

$$\left[\sum_{i=1}^{k} (F_{2n_i,2n_i,\alpha/2k} \cdot \bar{y}_i/\bar{t}_i),\ \sum_{i=1}^{k} (F_{2n_i,2n_i,1-\alpha/2k} \cdot \bar{y}_i/\bar{t}_i)\right].$$

Extending this approach gives a $(1-\alpha, 100\%$ confidence interval for MTUT denoted as $(L_{\alpha/2}, U_{1-\alpha/2})$. Expressions for $L_{\alpha/2}$ and $U_{1-\alpha/2}$ are given in Appendix B. Decision rules for the following tests: (a) H_o: MTUT $\leq C_o$ versus H_A: MTUT $> C_o$, (b) H_o: MTUT $\geq C_o$ versus H_A: MTUT $< C_o$, and (c) H_o: MTUT $= C_o$ versus H_A: MTUT $\neq C_o$ can be derived using the expressions for $L_{\alpha/2}$ and $U_{1-\alpha/2}$. Take for example the test H_0: MTUT $\geq C_0$ versus H_A: MTUT $< C_0$. A test with type I error probability of α would correspond to rejection of H_0 if $U_{1-\alpha}$ is less than C_0, otherwise fail to reject the null hypothesis. Procedures for the other two testing scenarios are straightforward.

The estimation and discrimination procedures outlined above are reasonably straightforward. The level of certainty for the interval estimates (at least $1-\alpha$) and error probabilities for the discrimination procedures are (at most α) very conservative. We turn now to estimation and testing for MTUT when conditional values of n_i, n_{ij}, and n_{ijl} are large.

The occurrence of an equipment failure produces values of five independent random variables. The random vector associated with a given failure is $\underset{\sim}{Z}_{ijlm} \equiv (t_{ijlm}, y_{ijlm}, x_{ijlm}, x^{*}_{ijlm}, x^{**}_{ijlm})$. The corresponding vector of parameters for this random vector can be defined as $\underset{\sim}{\Theta}_{ijl} = (\lambda_{ijl}, Mct_i, Mct_i^{*}, Mct_{ij}, Mct_{ijl})$. The likelihood function associated with the occurrence of any particular equipment failure is seen to be

$$f(\underset{\sim}{Z}_{ijlm}; \underset{\sim}{\Theta}_{ijl}) \equiv \lambda_{ijl}\,(1/Mct_i)(1/Mct_i^{*}) \cdot (1/Mct_{ij}) \cdot (1/Mct_{ijl})$$

$$\cdot \exp\left[-(\lambda_{ijl} \cdot t_{ijlm} + y_{ijlm}/Mct_i + x_{ijlm}/Mct_i^{*} + x^{*}_{ijlm}/Mct_{ij} + x^{**}_{ijlm}/Mct_{ijl})\right].$$

Thus, the likelihood for the conditional values of n_i, n_{ij}, and n_{ijl} failures is

$$L(\underset{\sim}{\Theta}) \equiv \prod_{i=1}^{k} \prod_{j=1}^{k_i} \prod_{l=1}^{k_{ij}} \prod_{m=1}^{n_{ijl}} f(\underset{\sim}{Z}_{ijlm};\ \underset{\sim}{\Theta}_{ijl})$$

where $\underset{\sim}{\Theta}$ is the vector of all parameter vectors $\underset{\sim}{\Theta}_{ijl}$. Actually, $L(\underset{\sim}{\Theta})$ depends on $\underset{\sim}{Z}$, where $\underset{\sim}{Z}$ is the vector of all random vectors $\underset{\sim}{Z}_{ijlm}$.

The value of the ijl^{th} portion of the vector $\underset{\sim}{\Theta}$, denoted by $\hat{\underset{\sim}{\Theta}}_{ijl}$, which maximizes $L(\underset{\sim}{\Theta})$, takes on the form $\hat{\underset{\sim}{\Theta}}_{ijl} = (\hat{\lambda}_{ijl}, \hat{Mct}_i, \hat{Mct}_i^*, \hat{Mct}_{ij}, \hat{Mct}_{ijl})$. This estimated vector can be thought of as the maximum likelihood estimator of $\underset{\sim}{\Theta}_{ijl}$. Expressions for these estimates and their respective large sample variances are given in Appendix C. Hence we see that

$$\hat{MTUT} \equiv \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} (1/\bar{t}_{ijl})(\bar{y}_i + \bar{x}_i + \bar{x}_{ij}^* + \bar{x}_{ijl}^{**})$$

is the maximum likelihood estimator of MTUT, being unbiased and normally distributed for large, conditional values of n_i, n_{ij}, and n_{ijl}. Using the estimated variance of the maximum likelihood estimator of MTUT described in Appendix C, the large sample $(1-\alpha)100\%$ confidence interval based on the maximum likelihood estimator is of the following form

$$\hat{MTUT} \pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{MTUT})}. \qquad (12)$$

This interval has an approximate confidence coefficient of $1-\alpha$ with the optimum property of being shorter, on the average, than intervals determined by any other estimator of MTUT.

Large-sample testing procedures for MTUT follow from expression [12].

The confidence coefficients and error probabilities associated with the approaches developed for large sample sizes are not conservative. Large-sample theory for maximum likelihood estimation guarantees the accuracy

of the confidence coefficients and error probabilities associated with the use of expression (12).

6. FUTURE RESEARCH DIRECTIONS

Two specific topics eligible for future development can be identified. The first topic is related to the measure ρ. How can ρ be estimated, given the total number of equipments being serviced and known forces of mortality for each equipment? What does the measure ρ represent when γ is known and all maintenance echelons are considered? These are two questions concerning the measure ρ and its application to maintenance support burden. The second area warranting future research efforts arises from the work presented in section 5. The $(1-\alpha)100\%$ interval estimate of MTUT for small values of n_{ijl} is very conservative. What is the "best" interval estimate of MTUT (best in the sense of shortest length, on the average, for a given level of confidence)? If an upper one-sided interval estimate is desired, what is the minimum upper one sided $(1-\alpha)100\%$ bound?

7. SUMMARY AND CONCLUSIONS

The concept of maintenance support burden is set forth in this paper as an important equipment characteristic to be considered in decision making. A comparison of standard maintainability/reliability measures with the mean overall maintenance time expressed as a percent of the average lifetime (MTUT) is discussed. Deficiencies of availability (A), MTTR, and mean queue length (ρ) as reflections of maintenance support burden are pointed out. A workable testing program for the evaluation of maintenance support burden of an equipment under consideration, useful in the demonstration phase of an equipment, is developed along with the respective estimation and discrimination procedures. Interval estimates and discrimination procedures

for a measure of maintenance support burden (MTUT) applicable in the assessment phase of an equipment is developed. These results motivate the importance of considering the proposed measure (MTUT) as part of a procurement decision making strategy as well as an assessment evaluation. Further, the significance of understanding the appropriate interpretation of standard maintainability/reliability measures in decision making situations will aid in a decrease of their misuse. In conclusion, the estimation and discrimination procedures for maintenance support burden, derived in this article, enhance the ability of reliability engineers to identify equipment types producing a minimum maintenance support burden (applicable in both the demonstration and assessment phase of development).

ACKNOWLEDGMENTS

This research was sponsored by the Air Force Office of Scientific Research, AFSC, United States Air Force, under Contract F49620-85-C-0013. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

I appreciate the constructive comments of the Editor and referees.

Appendix A

Expressions for $\hat{Var}\,(\hat{MTUT})$ in both the demonstration and assessment stages described in section 4 are derived below.

Demonstration Stage

Defining $w_{ijlm} = (y_{ijlm} + x_{ijlm} + x^{*}_{ijlm} + x^{**}_{ijlm})$

We have $$\bar{w} = \{\sum_{i=1}^{k}\sum_{j=1}^{k_i}\sum_{l=1}^{k_{ij}}\sum_{m=1}^{n_{ijl}} w_{ijlm}\}/30$$

and

$$s_w^2 = \{\sum_{i=1}^{k}\sum_{j=1}^{k_i}\sum_{l=1}^{k_{ij}}\sum_{m=1}^{n_{ijl}} (w_{ijlm} - \bar{w})^2\}/29$$

for $n_{ijl} > 0$.

Thus,

$$\hat{Var}\,\hat{MTUT} = \lambda_T^2 \cdot [s_w^2/30].$$

Assessment Stage

We define for large n_{ijl} the following:

$$\bar{y}_i = \{\sum_{j=1}^{k_i}\sum_{l=1}^{k_{ij}}\sum_{m=1}^{n_{ijl}} y_{ijlm}\}/n_i$$

$$\bar{x}_i = \{\sum_{j=1}^{k_i}\sum_{l=1}^{k_{ij}}\sum_{m=1}^{n_{ijl}} x_{ijlm}\}/n_i$$

$$\bar{x}^{*}_{ij} = \{ \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} x^{*}_{ijlm} \}/n_{ij}$$

$$\bar{x}^{**}_{ijl} = \{ \sum_{m=1}^{n_{ijl}} x^{**}_{ijlm} \}/n_{ijl},$$

where

$$s^2_i = \{ \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} (y_{ijlm} - \bar{y}_i)^2 \}/(n_i - 1)$$

$$s^{*2}_i = \{ \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} (x_{ijlm} - \bar{x}_i)^2 \}/(n_i - 1)$$

$$s^2_{ij} = \{ \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} (x^{*}_{ijlm} - \bar{x}^{*}_{ij})^2 \}/(n_{ij} - 1)$$

$$s^2_{ijl} = \{ \sum_{m=1}^{n_{ijl}} (x^{**}_{ijlm} - \bar{x}^{**}_{ijl})^2 \}/(n_{ijl} - 1).$$

Thus,

$$\hat{\text{Var}}\ \hat{\text{MTUT}} = \sum_{i=1}^{k} \lambda_i^2 (s_i^2 + s^{*2}_i)/n_i + \sum_{i=1}^{k} \sum_{j=1}^{k_i} \lambda_{ij}^2 \cdot s^2_{ij}/n_{ij}$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \lambda_{ijl}^2 \, s^2_{ijl}/n_{ijl}.$$

Appendix B

This appendix gives the upper and lower confidence bounds for MTUT with small conditional values of n_i, n_{ij}, and n_{ijl} (λ_{ijl}'s are unknown). Defining $g = \sum_{i=1}^{k} \sum_{j=1}^{k_i} k_{ij}$, $d = \sum_{i=1}^{k} k_i$, and $\delta = 2k + g + d$ the expressions for $L_{\alpha/2}$ and $U_{1-\alpha/2}$ are given by the following:

$$L_{\alpha/2} = \sum_{i=1}^{k} (F_{2n_i, 2n_i; \alpha/2\delta})(\bar{y}_i + \bar{x}_i)/\bar{t}_i$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} (F_{2n_{ij}, 2n_{ij}; \alpha/2\delta} \cdot \bar{x}_{ij}^{*})/\bar{t}_{ij}$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} (F_{2n_{ijl}, 2n_{ijl}; \alpha/2\delta} \cdot \bar{x}_{ijl}^{**})/\bar{t}_{ijl}.$$

$$U_{1-\alpha/2} = \sum_{i=1}^{k} (F_{2n_i, 2n_i; 1-\alpha/2\delta}) \cdot (\bar{y}_i + \bar{x}_i)/\bar{t}_i$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} (F_{2n_i, 2n_i, 1-\alpha/2\delta} \cdot \bar{x}_{ij}^{*})/\bar{t}_{ij}$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} (F_{2n_{ijl}, 2n_{ijl}; 1-\alpha/2\delta} \cdot \bar{x}_{ijl}^{**})/\bar{t}_{ijl}.$$

Note: $\bar{t}_i = \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} t_{ijlm}/n_i$

$$\bar{t}_{ij} = \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} t_{ijlm}/n_{ij}$$

$$\bar{t}_{ijl} = \sum_{m=1}^{n_{ijl}} t_{ijlm}/n_{ijl}.$$

Appendix C

Maximum likelihood estimates of the parameters in the vector $\underset{\sim}{\Theta}_{ijl}$ are given below. The variance of $\hat{MTUT}$ is also developed. Substituting the respective maximum likelihood estimators of Mct_i, Mct_i^*, Mct_{ij}, Mct_{ijl} and λ_{ijl} into the functional form for $Var(MTUT)$ gives $\hat{Var}(\hat{MTUT})$. This substitution does not appreciably affect the accuracy of the maximum likelihood estimate of $\hat{Var}(MTUT)$ according to Mood, Graybill, and Boes [3]. The resulting estimated variance is used in the interval estimation of MTUT in section 5. Because of the independence and distribution assumptions associated with the repair and life times under consideration, we see that

$$\hat{\lambda}_{ijl} \equiv 1/\bar{t}_{ijl},\ \bar{t}_{ijl} = \sum_{m=1}^{n_{ijl}} t_{ijlm}/n_{ijl}$$

$$\hat{Mct}_i \equiv \bar{y}_i = \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} y_{ijlm}/n_i$$

$$\hat{Mct}_i^* \equiv \bar{x}_i = \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} x_{ijlm}/n_i$$

$$\hat{Mct}_{ij} \equiv \bar{x}_{ij}^* = \sum_{l=1}^{k_{ij}} \sum_{m=1}^{n_{ijl}} x_{ijlm}^*/n_{ij}$$

$$\hat{Mct}_{ijl} \equiv \bar{x}_{ijl}^{**} = \sum_{m=1}^{n_{ijl}} x_{ijlm}^{**}/n_{ijl}$$

are the maximum likelihood estimators of the respective parameters included in $\underset{\sim}{\Theta}_{ijl}$. According to Mood, Graybill and Boes [3], $\hat{\underset{\sim}{\Theta}}_{ijl}$ has a mean vector $\underset{\sim}{\Theta}_{ijl}$ and a variance-covariance matrix whose off-diagonal elements are zero and diagonal elements being

$$\hat{Var}\ \hat{\lambda}_{ijl} \equiv \lambda_{ijl}^2/n_{ijl}$$

$$\text{Var } \hat{\text{Mct}}_i \equiv (\text{Mct}_i)^2/n_i$$

$$\text{Var } \hat{\text{Mct}}_i^* \equiv (\text{Mct}_i^*)^2/n_i$$

$$\text{Var } \hat{\text{Mct}}_{ij} \equiv (\text{Mct}_{ij})^2/n_{ij}$$

$$\text{Var } \hat{\text{Mct}}_{ijl} \equiv (\text{Mct}_{ijl})^2/n_{ijl}$$

when n_i, n_{ij}, and n_{ijl} are large. Hence, we see that the maximum likelihood estimator $\hat{\text{MTUT}} \equiv \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} (1/\bar{t}_{ijl})(\bar{y}_i + \bar{x}_i + \bar{x}_{ij}^* + \bar{x}_{ijl}^{**})$ is unbiased for MTUT, normally distributed, and

$$\text{Var } \hat{\text{MTUT}} \equiv \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{l=1}^{k_{ij}} [(\text{Mct}_i + \text{Mct}_i^* + \text{Mct}_{ij} + \text{Mct}_{ijl})^2 \lambda_{ijl}^2/n_{ijl}$$

$$+ [\lambda_{ijl}^2/n_{ijl} + (\lambda_{ijl})^2] \cdot [\text{Mct}_i^2/n_i + (\text{Mct}_i^*)^2/n_i + \text{Mct}_{ij}^2/n_{ij}$$

$$+ \text{Mct}_{ijl}^2/n_{ijl}]]$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{\substack{m=1 \\ m \neq n}}^{k_{ij}} \sum_{n=1}^{k_{ij}} \lambda_{ijm} \lambda_{ijn} (\text{Mct}_i)^2/n_i$$

$$+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq p}}^{k_i} \sum_{p=1}^{k_i} \sum_{m=1}^{k_{ij}} \sum_{n=1}^{k_{ij}} \lambda_{ijm} \lambda_{ipn} (\text{Mct}_i)^2/n_i$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{\substack{m=1 \\ m \neq n}}^{k_{ij}} \sum_{n=1}^{k_{ij}} \lambda_{ijm} \lambda_{ijn} (\overset{*}{Mct}_i)^2/n_i$$

$$+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq p}}^{k_i} \sum_{p=1}^{k_i} \sum_{m=1}^{k_{ij}} \sum_{n=1}^{k_{ip}} \lambda_{ijm} \lambda_{ijn} (\overset{*}{Mct}_i)^2/n_i$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k_i} \sum_{\substack{m=1 \\ m \neq n}}^{k_{ij}} \sum_{n=1}^{k_{ij}} \lambda_{ijm} \lambda_{ijn} (Mct_{ij})^2/n_{ij}.$$

REFERENCES

[1] Ascher, H., "Mil-Std-781C: A Vicious Circle," IEEE Trans. Reliability, vol. R-36, no. 4, 1987, Oct., pp. 397-402.

[2] Ascher, H., H. Feingold, Repairable Systems Reliability, New York, New York, Marcel Dekker Inc., 1984, p. 150.

[3] Boes, D.C., F.A. Graybill, and A.M. Mood, Introduction to the Theory of Statistics, 3rd edition, McGraw-Hill, Inc.: New York, 1974, pp. 358-360, 393-396.

[4] Klion, J., "Specifying Maintainability - A New Approach," Proc. Annual Reliability and Maintainability Symposium, Las Vegas, Nevada, January 1986, pp. 338-343.

[5] Parzen, Emanuel, Stochastic Processes, San Francisco, California, Holden-Day, 1962, pp. 147, 148, 281, 282.

[6] Thompson, W.A., "On the Foundations of Reliability," Technometrics, vol. 23, 1981, pp. 1-13.

FINAL REPORT NUMBER 44
REPORT NOT AVAILABLE AT THIS TIME
Dr. Louis Johnson
760-7MG-050

FINAL REPORT

on

SIGNED-DIGIT NUMBER SYSTEM FOR OPTICAL ADAPTIVE PROCESSING

by

P.A. Ramamoothy
Associate Professor of Electrical & Computer Engineering
University of Cincinnati
Mail Location #30
Cincinnati, Ohio 45221

Submitted to

AFOSR Research Initiation Program
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

ABSTRACT

There is great deal of interest in the Air Force in adaptive processing in beam forming networks to make signal receivers less susceptible to degradation in signal-to-noise ratio caused by undesired noise signals and interference from electronic countermeasure systems. However, Adaptive beam forming (ABF) is computationally intensive and becomes unmanageable as the number of elements in the beam former is increased. This has led to interest in optical implementation as it can provide 100 to 1000 times the throughput rate that is difficult to achieve using electronic implementation. In this work, we have studied signed-digit number system (SDNS) and modified signed-digit number system and their application in optical implementation. In particular, we have studied signed-digit numbers with radix greater than two and proposed methods to represent both integer and floating point numbers. We have also shown the implementation of basic units such as adders, and multipliers in SDNS using optical elements. Through this work we have shown that the SDNS has certain properties that make it possible to fully utilize important properties of optics such as massive parallelism and also to overcome drawbacks of other number systems such as binary or residue.

1. INTRODUCTION

Adaptive beam forming (ABF) is used extensively in defense radar and satellite communication systems to track desired signals and provide protection against jamming and other interference signals [1-3]. ABF provides the ability to sense automatically the presence of interference noise signals and to suppress them while simultaneously enhancing desired signal reception. This is achieved with very little or no prior knowledge of the signal/interference environment. However, ABF is computationally intensive and can tax the capabilities of even super-computers. This computational problem becomes more critical as more elements are added to adaptive beam forming network to counter the growing and/or expected threat. For example, recent specifications from AF/RADC calls for 256 to >1000 elements, processing bandwidth of 50 MHz to 1 GHz, and weight update rates of 100 KHz to 2 MHz as far term goals [4]. Such goals cannot be realistically achieved even using massive parallelism in electronic implementation. Hence there is a great push towards optical processing to perform such high speed adaptive processing tasks.

Optical implementation (OI) is considered to be superior to electronic implementation (EI) since it offers massive parallelism, dynamic reconfigureability and high local and global interconnectivity (made possible by free-space interconnection). However, to harness, even

partially, the benefits offered by optics, entirely new thinking and approaches have to be developed. Thus, instead of attempting to emulate techniques that are used in EI, one will have to develop new techniques. Number system selection is a good example.

Number systems such as binary (which does not make use of parallelism in optics) or residue number system (which allows only integer number representation and does not make use of the high interconnectivity property of optics) that are popular in EI cannot be very effective in OI. Hence, in this study, we have taken a look at the signed-digit number system and compare the merits and shortcomings with other number systems. The signed-digit number system which was proposed almost 25 years ago [5], did not ever catch on in EI and there is no follow-on work and/or publications. Recently, there has been some interest in a special case of signed-digit number system, that of modified signed-digit (MSD) number system where radix equals two, and its use in optical implementation has been demonstrated [6-9], the last three articles by the principal investigator. Since it has been shown that MSD number system makes use of the parallelism offered by OI, signed-digit has the same property, and hence one can make use of the property of OI to signed-digit number system.

SPECIAL OBJECTIVES

The objective of this research work is to evaluate the usefulness of signed-digit number system with radix r (r>2) for optical processors in general and for optical adaptive processors, in particular. The work involved the study of:

a) merits and drawbacks of such a number system,

b) representation of both integer and floating-point numbers,

c) algorithms for basic functions such as addition, barrel-shifting (multi-digit shifting) and multiplication,

d) suitable architectures for implementing those algorithms using optical elements, and

e) Their use in optical adaptive beam forming.

From the results presented in subsequent sections, it can be seen that the signed-digit number system have certain properties that make them strong candidate for optical processing.

2. SIGNED-DIGIT NUMBER SYSTEM (SDNS)

Signed-digit number system is a redundant number system that offers certain advantages, for example, SDNS limits carry propagation to one position to the left during the operations of addition and subtraction in digital computers. Carry-propagation chains are eliminated by the use of redundant representations for the operands. The digits of a signed-digit representation individually assume both

positive and negative integer values and contain the sign information for the number; no separate sign information for the given number is necessary. In general, a signed-digit number is represented by n+m+1 digits where each digit z_i is such that $|z_i| \leq r-1$ for radix r and has the algebraic value

$$Z = \sum_{i=-n}^{m} z_i r^{-i} \qquad (1)$$

where r>2. In this representation the algebraic value Z=0 has a unique representation, i.e. when all the digits are zero. Sign of the algebraic value of Z can be inferred from the sign of its most significant digit. Also, the representation for -Z can be obtained from the representation of Z by changing signs for all the digits. In the case of r=2 the representation is known as the Modified Signed Digit (MSD) representation. This representation has been used in several optical systems [6]. In the MSD number representation addition is performed in three stages instead of two stages for the normal SD representation.

3. SIGNED-DIGIT ARITHMETIC

Given the class of signed-digit representations described by (1) this section describes basic operations on these numbers.

3.1. Addition/Subtraction

The addition of two signed-digit numbers is performed in parallel in two successive steps. First, an outgoing transfer digit t_{i+1} and intermediate sum digit w_i are produced and then the sum digit s_i is formed as shown in Fig. 1. In this figure, five digit operands are assumed and the structure is independent of the radix r (r>2).

$$z_i + y_i = rt_i+1 + w_i \tag{2}$$

$$s_i = w_i + t_i \tag{3}$$

Parallel addition without carry propagation in SD arithmetic is achieved by imposing restrictions on values of t_i and w_i and these are $|t_i| \leq 1$ and $|w_i| \leq r-2$ respectively. Thus, given the allowed values for w_i as the sequence $w_{min},\ldots,-1, 0,1,\ldots,w_{max}$, the rules for finding w_i, t_i, and s_i are as follows:

$$w_i = (z_i + y_i) - rt_{i+1} \tag{4}$$

where

$$t_{i+1} = \begin{cases} 0 & \text{if } w_{nim} \leq z_i + y_i \leq w_{max} \\ 1 & \text{if } z_i + y_i \geq w_{max} \\ -1 & \text{if } z_i + y_i \leq w_{nim} \end{cases}$$

and then

$$s_i = w_i + t_i \tag{5}$$

To perform subtraction the property of deriving the representation of -Z from Z can be used i.e. change the sign of all the digits before feeding to the adder and it will perform a subtraction.

3.2. Multi-digit Adder (MDA)

The redundancy in SDNS can be exploited to add (sub) more than two numbers simultaneously. In this section, we describe this basic concept and present the adder architecture for such an addition.

The advantage of using large radix and large digit set is to maximally employ redundancy of SDNS. We present here a way to achieve the multi-digit adder. The large sets of digit allow the possibilities of summing up to r/2 numbers at the same time, r is the selected radix. For example, let r=8, if appropriate "carry" and "sum" look-up tables are chosen, then up to 4 SDNS numbers (octal) can be added together simultaneously.

Taking r=8 as the example, the maximum value of adding 4 octal single digits together is 34_8. The minimum value of adding 4 octal digits is $\bar{3}\bar{4}_8$. If the maximum digit set is selected, then we can set the ranges for the tables "carry" and "sum". The maximum digit set for r=8 is [-7,-6,-5,-4, -3,-2,-1,0,1,2,3,4,5,6,7]. The range of table "carry" should be set as ≥ 3 to ≤ 3; and the range of "sum" table should be set as ≥ 4 to ≤ 4. Then the SD adder may be constructed as in Fig. 2.

The adder proposed here will increase the speed of addition by 4 times than the adder using SDA's for both stages. Generally, if radix is r and the maximum digit set is used, the time saving will be a factor of r/2.

The most difficulty problem associated with the multi-digit SDNS adder is the extensively increased look-up tables

needed, for the example of r=8, we would need $2x8^4$ table entries for stage one MDA's, and 450 entries for stage two blocks. One way to solve this problem is to construct the look-up tables such that only minimum table entries are needed, since most of table entries contain the same numbers. For r=8 example, a total of 15 different entries are needed. Another method to solve this problem is to use conventional two digits adder to construct MDA's. Fig. 3 shows the internal construction of stage one MDA block. There FA's are digits adders. Then w_i' and c_{i+1}' are checked by the limits of ranges set up earlier. For r=8, if $w_i'>4$, let w_i to be the complement of w_i' and c_{i+1} to be $c_{i+1}'+1$. If $w_i'\leq 4$, it is within the range, then let w_i to be w_i', and c_{i+1} to be c_{i+1}'. And then these two lines are feed into stage two blocks.

3.3. Multiplication

Multiplication can be performed in several different ways - sequential add & shift, array multiplier (with some modifications), or by generating partial products and then summing them. Though the array multiplier requires only local interconnections, the delay depends on the number of digits in the operands. Fast multiplication can be performed in the signed-digit representation by first generating all the partial products in constant time independent of the number of digits and then summing these products using a binary tree. This kind of scheme was proposed by Takagi [10]. For a radix 4 (digit set [-3,-2,

-1,0,1,2,3]) it can be seen that the partial product generator is the configuration shown in Fig. 4. The radix r=4 has been chosen here because of simplicity of the partial product generator. The appropriate unit is activated by the control signal depending on the multiplier digit. As a shift in radix 4 amounts to a multiplication of the number by 4, multiplication by 3 can be performed by doing a left shift and then subtracting the original number once. NOP is the no-operation phase where an input is passed onto the output.

The multiplier configuration for n=4 is shown in Fig. 5. In this figure $X_3X_2X_1X_0$ are the multiplicand digits and $Y_3Y_2Y_1Y_0$ are the multiplier digits. The signed-digit adder (SDA) has to have an appropriately larger width to incorporate the shift in the addition of the partial products, but this does not introduce any more delay as addition is performed in constant time. Thus the delay of the multiplier is $\log_2 n$ (depth of the binary tree) where n is the number of digits (radix 4 in this case).

It is also possible to perform the addition of the partial products in a higher radix signed-digit representation without generating a carry. Thus, several partial products can be added together simultaneously to reduce the depth of the tree. It is particularly simple to go back to the original radix if the new radix is a higher power of the original radix. To go back to the original radix, one has to only regroup the binary sequence

representing the SD number. Therefore in Fig. 5, the addition of the 4 partial products (with r=4) could have been performed by a radix 16 adder. It would generate no carry. The delay of the multiplier would be constant if only working in the higher radix was easier, but it is observed that for higher radix the number of terms required to implement the adder increase drastically.

3.4. Multiplication using MDA's

In this section, we present the architecture of a multiplier which is constructed by multi-digit adders (MDA). One of the disadvantages of using large radix or large digit set is that the algorithms of SDNS multiplication are very complicated, and array type multiplier are difficult to be built. But if the MDA's are used in design of multipliers, a lot of work will be saved.

Since MDA can add r/2 digits at the same time, for each multiplier digit we can either perform addition only once (if multiplier digit $\leq$ r/2) or twice (if multiplier digit > r/2). Fig. 6 is an example of a partial product generator (PPG) for a SDNS multiplier. The Comparator block is to decide performing one or two times of MDA according to multiplier digit Y_i. The multiplier using these PPG's is shown in Fig. 7, where r=8, multiplicand is $X_nX_{n-1}...X_1X_0$, and multiplier is $Y_nY_{n-1}...Y_1Y_0$. The delay of this scheme is $\log_{r/2}n$.

3.5. Algebraic Comparison

Algebraic comparison can be performed as it is done in the conventional binary representation schemes, by subtracting one number from the other. As subtraction is a fast operation in signed-digit arithmetic, algebraic comparison can be performed faster than is possible in the binary representations. Also this can be contrasted with the fact that algebraic comparison in residue arithmetic is impossible.

3.6. Division

Division in the signed-digit representation is slightly more complicated as the quotient bits for radix 4 belong to the digit set [-3,-2,-1,0,1,2,3]. Sign of the quotient digit is negative if the signs of the most significant digits of the dividend at that stage and the divider are different. To speed up the division one can use a set of comparators (subtractors) and multipliers to obtain the quotient digit and a subtractor to take the difference. As division is not required too often in most operations this is not expected to be a drawback. The fact that division is possible in the SD representation can be contrasted with the fact that it cannot be performed in residue arithmetic and where it is not closed (being an integer representation) under division.

3.7. Conversion

It is advantageous to use a radix > 2 which is a power of 2 as this allows easy conversion from the sign-magnitude binary form to the signed-digit form by grouping together

$\log_2 r$ bits and forming the digit. If the sign is negative then change the signs of all the digits to negative.

To convert from the signed-digit representation to the binary representation it is again of advantage to use a radix which is a power of 2. Form two separate numbers - one containing the positive digits and the other containing the negative digits. Then convert each of these digit representations to binary - each digit contributing $\log_2 r$ bits and subtract the negative binary set from the positive binary set to yield the equivalent binary representation.

Other sequential methods are also possible for performing the conversion from signed-digit to conventional binary. By inspecting the SD number (say in radix 4) from the most significant digit and analyzing two digits at a time one can eliminate all other digits except the 0,1 set (this is possible because SD representation is redundant and not unique). Several passes may be required but at the end we have the binary representation of the number.

4. SELECTION OF PROPER RADIX AND PROPER DIGIT SET

For SDNS numbers, any radix > 2 can be chosen, but we need to identify the one with most advantage offered by such a number system. And similarly, for a given radix we have several digit sets to choose from. In this section, we will discuss the selection of radix and digit set so that it will maximally take the advantage of redundancy of SDNS.

4.1. Background

The SDNS is a redundant number system with a radix greater than two. The SDNS with a radix r has the digit set:

$$D_{sd} = [-n,-(n-1),\ldots,-1,0,1,\ldots(n-1),n]$$

where the value of n is given by

$$n_{min} \leq n \leq n_{max} \quad \text{and}$$

$n_{min} = (r+1)/2$ if r is odd, $n_{min} = r/2 + 1$ if r is even; $n_{max} = r-1$. Therefore, for SDNS with radix r, we can have digit sets:

$$D_{sd\ min} = [-n_{min},-(n_{min}-1),\ldots,-1,0,1,\ldots,n_{min}-1,n_{min}]$$

. . .

$$D_{sd\ max} = [-n_{max},-(n_{max}-1),\ldots,-1,0,1,\ldots,n_{max}-1,n_{max}].$$

For minimum set, there are $2n_{min}+1$ different representations (digits), while there are 2r-1 digits for the maximum set. For any value, say N, is represented in SDNS as:

$$(N)_r = D_{sd\#p}r^p + D_{sd\#p-1}r^{p-1} +\ldots+ D_{sd\#1}r + D_{sd\#0}$$

It is true that decimal (r=10) is the number system with which everyone is familiar and comfortable, but the difficulty of converting a decimal number into binary with which computers operate makes it out of our consideration. Our guideline of the selection of proper radix is that which radix makes addition and multiplication easier. Of course, smaller radices make adder and multiplier simpler, but larger radices offer larger redundancy which means more effective algorithms might be used to construct the adder and multiplier. Much the same for the selection of proper

digit set. Smaller digit sets lack of redundancy but make things simpler; larger digit sets have more redundancy but need more hardware to configure the adder and multiplier.

4.2. Comparison of Different Radices

Theoretically we can have large number of radices as the base of SDNS. But implementation becomes complex as r getting large. Therefore we only consider the cases of $r \leq 16$.

The list below is a summary of different radices, their ranges, n_{min} and n_{max}, number of digits in sets, and number of sets for each radix. Binary MSD listed in the list is for a comparison:

r	# of sets	n_{min}	n_{max}	# of digits$_{min}$	# of digits$_{max}$
2	1	1	1	3	3
3	1	2	2	5	5
4	1	3	3	7	7
5	2	3	4	7	9
6	2	4	5	9	11
7	3	4	6	9	13
8	3	5	7	11	15
10	4	6	9	13	19
16	7	9	15	19	31

From the list, one can see r=3 or 4, only one digit set is there. The large radices have more digit sets but number of digits also are big, this will need more representations also. From the point of view of simplifying interfacing with electronic computers and the internal architecture of adder and multiplier, radix r which is a power of 2 will be the best candidate. Selection of radix as a power of 2 will

also make it possible for easy conversion between signed-digit and conventional binary digit.

Since look-up table method is proposed to be used in the algorithms of addition of SDNS numbers, then the number of entries on those tables also need to be considered when we select a proper radix or digit set. The following is a list which gives the number of table entries for different radices:

r	n_{min}	n_{max}
3	50	50
4	98	98
5	98	162
6	162	242
7	162	338
8	242	450
10	338	722
16	722	1922

There are two tables for each radix, one for sum produced by add two SDNS digits, one for carry. From the list above, we can see that selecting the smaller set will make look-up table smaller. It might be important for considering selecting the larger set if the memory capacity is critical.

5. MATRIX OPERATIONS

Primitive arithmetic operations defined in the previous sections can be used to implement various operations on matrices. In this section systolic arrays proposed by Mead & Conway [11] are used to perform matrix-vector and matrix-matrix multiplication, and solution to the problem of least

square minimization by Q-R decomposition [12]. As the details are available elsewhere only the salient points of these operations are described here.

5.1. Matrix-vector Multiplication

The inner-product step processor performs the operation C <-- C + A x B. This processor is used as the basic cell to perform the matrix-vector multiplication. As the signed-digit multiplier is implemented by adding together the partial products it is quite easy to implement the inner product step processor operation which is another stage of addition. At each clock cycle each of these processors receive new inputs for A, B, and C, perform the computation and output the result in the next clock cycle. To multiply an n x n matrix $A = (a_{ij})$ with an n x 1 vector $\underline{x}^T = (x_1 \; x_2 \ldots x_n)$ and obtain the product, an n x 1 result $\underline{y}^T = (y_1 \; y_2 \ldots y_n)$. The following recurrences are used:

$$y_i^1 = 0$$
$$y_i^{k+1} = y_i^k + a_{ik}x_k$$
$$y_i = y_i^{n+1}$$

The matrix-vector multiplier is shown in Fig. 8. All n components of $\underline{y}$ are computed in 4n - 1 clock cycles. As only 1/2 the number of processors are active at any clock cycle it is possible to use only n processors to perform the same operation

5.2. Matrix-matrix Multiplication

The same inner-product processor used for matrix-vector multiplication can be used for matrix-matrix multiplication.

The interconnections are better shown by using a hexagonal geometry for the same processor. The inner working of the inner-product processor is the same here and data moves in the same fashion as that for matrix-vector multiplication. Two n x n matrices $A = (a_{ij})$ and $B = (b_{ij})$ are multiplied to obtain an n x n matrix $C = (c_{ij})$ by the following recurrences:

$$\begin{aligned} c_{ij}^{1} &= 0 \\ c_{ij}^{k+1} &= c_{ij}^{k} + a_{ik}b_{kj} \\ c_{ij} &= c_{ij}^{n+1} \end{aligned}$$

n^2 hex-connected processors and it takes 4n clock cycles to perform the multiplication of these matrices. It is also possible to use $n^2/3$ processors as only one out of 3 processors is active at any time.

5.3. Least Square Minimization

Adaptive combiners for adaptive signal processing can be formulated in terms of least squares minimization. Inputs to the combiner take the form of a desired signal y(i) and N-1 auxiliary signals $\underline{x(i)}$, and the complex weight vector $\underline{w}$ is adjusted so as to minimize the power of the combined output signal

$$e(i) = y(i) + \underline{x}^T(i)\underline{w}, \text{ for } 1 \leq i \leq n \qquad (6)$$

By doing this, interference nulls can be created in other directions besides the direction of interest. The least square weight vector at time t_n is given by

$$X^H(n)X(n)\underline{w}(n) = X^H(n)\underline{y}(n), \qquad (7)$$

where X(n) is the n x N-1 data matrix, $\underline{y}(n)$ is an n x 1 vector of desired signal y(i), and X^H denotes the Hermitian of the matrix X.

Solving this equation for $\underline{w}(n)$ is complicated and the matrix $X^H(n)X(n)$ may have a very small determinant (ill-conditioned) and thus give large variations in solution for $\underline{w}(n)$. The Q-R decomposition technique using Givens rotations is more appropriate [13,14]. In this technique an n x n unitary (such that $Q^HQ = I$) matrix Q is found such that

$$Q(n)X(n) = (\ R(n)\ \ 0)^T \qquad (8)$$

and R(n) is an N-1 x N-1 upper triangular matrix. Gentleman and Kung have shown how this can be implemented using a triangular systolic array using three types of basic cells. The upper triangular matrix R(n) is recursively generated where each row of cells in the array performs a basic Givens rotation between a row of the stored triangular matrix and a vector of data. Once the triangularization has been performed a linear systolic array computes the least-squares weight vector by backward substitution. In the adaptive antenna application we are interested only in the beamformed signal and not in explicitly computing the weight vector. Thus the Q-R decomposition can be modified to directly compute the residual at each stage and the linear systolic array is no longer necessary. The configuration for this array and the cells are shown in Fig. 9.

6. IMPLEMENTATION

For comparison purposes with residue arithmetic implementation techniques of the least square solution, the Westinghouse design for adaptive phased array radars was chosen [15]. Assuming the same bound of 1.9×10^{15} it was found that 25 digits of radix 4 are required. In residue arithmetic out of necessity, a fixed point representation is used while in SD a floating point representation can be used. Position encoding and look-up tables as used in residue arithmetic operations are assumed. Look-up tables (LUT) can be created for performing the operation of addition and complementing in one delay. Thus the adder would require one delay, the subtractor two delays, partial-product generator three delays (assuming the shift operation requires no delays), multiplier $3 + \log_2 n$ delays (n=25 here so 8 delays), and the divider 3 + n =28 delays. If MDA's are used for adder or multiplier, then the number of delays are the same since r = 4. For large radix r, the delays of addition and multiplication will be reduced if MDA's are used. Each boundary cell in the triangular systolic array performs four sequential operations - two multiplications, addition and a division, thus requiring a total of 45 delays. The cells in the triangular array requires two multiplications, one addition and one subtraction, thus requiring 19 delays. As it takes more than one delay to generate the outputs of the cells, the clock width would be

that of the more limiting value, i.e. 45 delays. Each output residual signal is generated from a data vector that was inputted 2(N-1) clock cycles earlier, that is a latency of 450 delays for a 6 x 6 matrix. After the first output has been generated it requires only 45 delays for each following output. These values are considerably higher than that required for the residue number representation (latency = 72) but with a few more clock cycles (1 for binary to signed-digit and the delay of the carry look ahead adder) the inputs and outputs are in the conventional representation. In comparison, for the residue number representation, much more complex conversion schemes (Chinese Remainder Theorem or Mixed Radix Conversion) are required to be implemented which would exceed the time required here. Also, as the operations have not been pipelined at the finest level in this implementation, there is much scope for improvement

7. CONCLUSION

In this research work the use of the signed-digit number system for fast processing has been investigated. It has been compared with the residue number system, which has been widely used to meet the speed requirements in the past. The signed-digit number system offers most of the advantages of parallel processing, without problems such as conversion to/from conventional binary representation, and hence is an

alternative worth considering. We have completed the architectures for designing SDNS adder and multiplier. The selection of proper radix and digit set would allow us to add up to r/2 numbers at the same time. And the architecture of using MDA's for designing multiplier is also presented in this report.

As pointed out before, it is necessary to use floating-point numbers to achieve the required range for adaptive beamforming and the SDNS can be extended to floating-point easily whereas it would not be straightforward in the residue number representation.

8. ACKNOWLEDGEMENT

This work was supported by Air Force Office of Scientific Research Grant #F760-7MG-015, and monitored by Universal Energy Systems, Inc. Parts of the results from this work have been presented at conferences of Optoelectronic Signal Processing for Phased-Array Antennas, 10-17 Jan. 1988; and Acoustics, Speech, and Signal Processing, 11-14 April 1988, and published in the proceedings of these conferences.

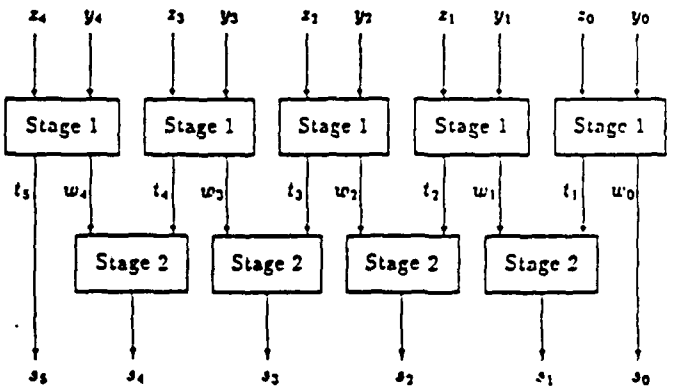
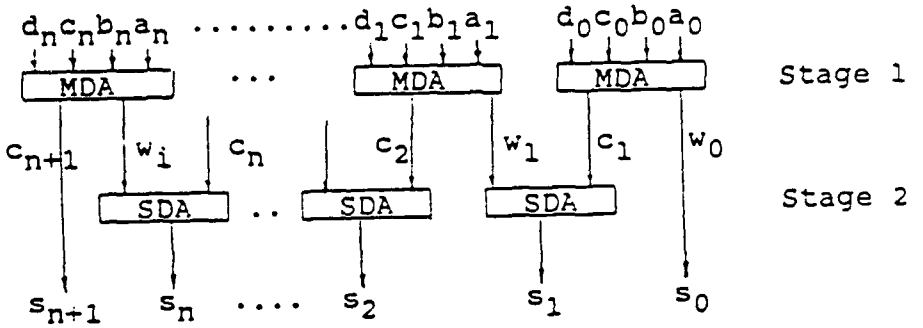


Fig. 1 Signed-digit Adder.



r=8, MDA - multi-digit adder of SDNS
SDA - single-digit adder of SDNS

Fig. 2 Multi-digit SD Adder.

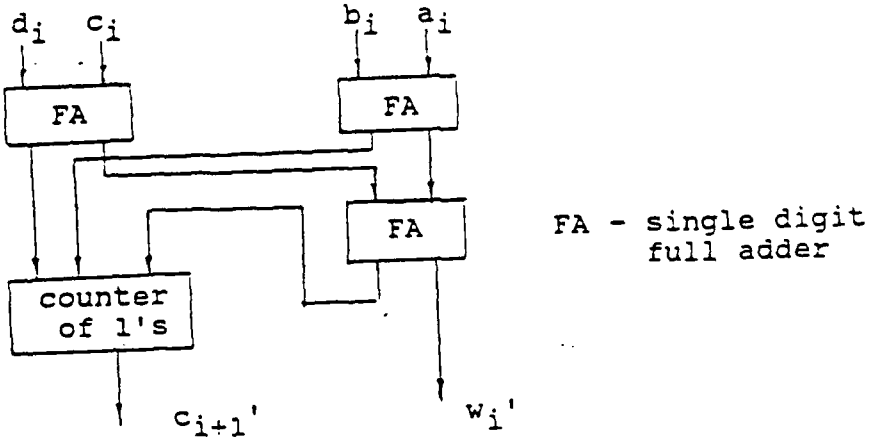


Fig. 3 Internal Architecture of Stage One MDA Blocks.

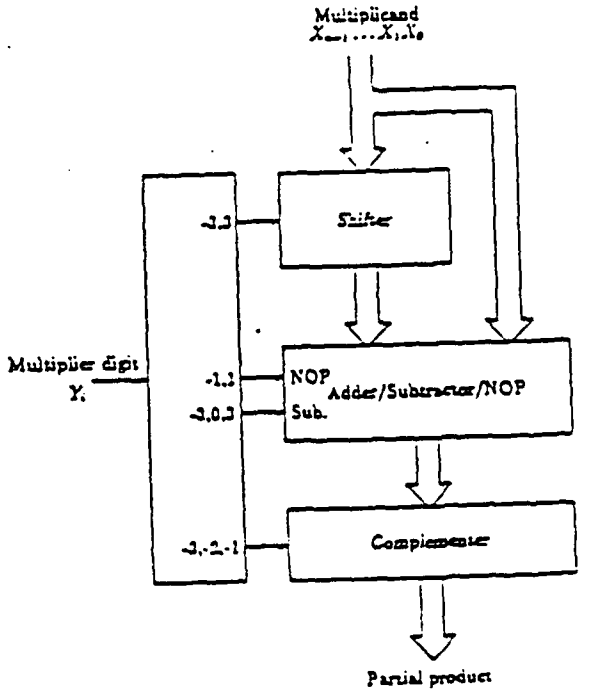


Fig. 4 Partial Product Generator for Radix 4.

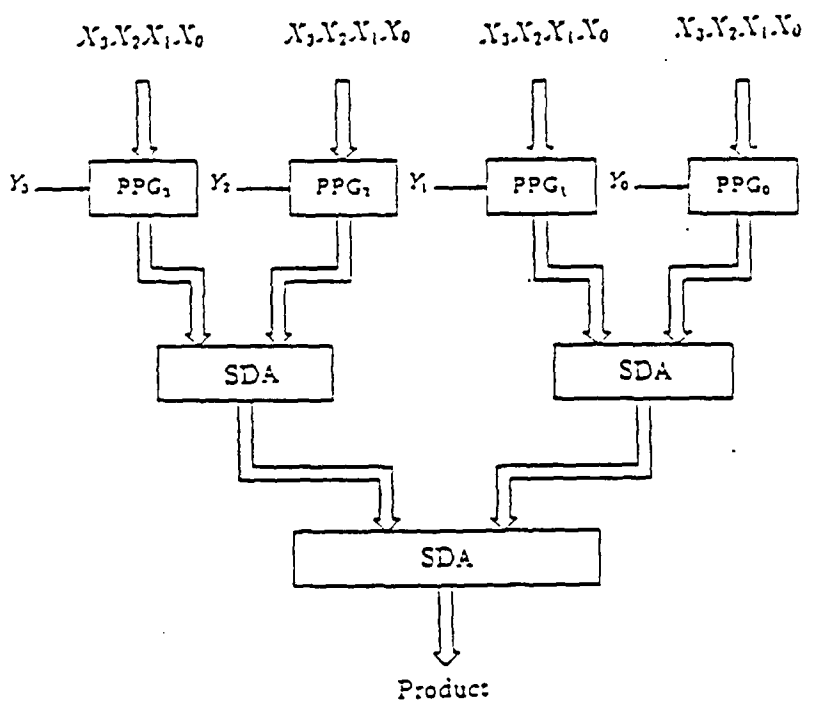


Fig. 5 Binary Tree for Addition of Partial Products (r=4).

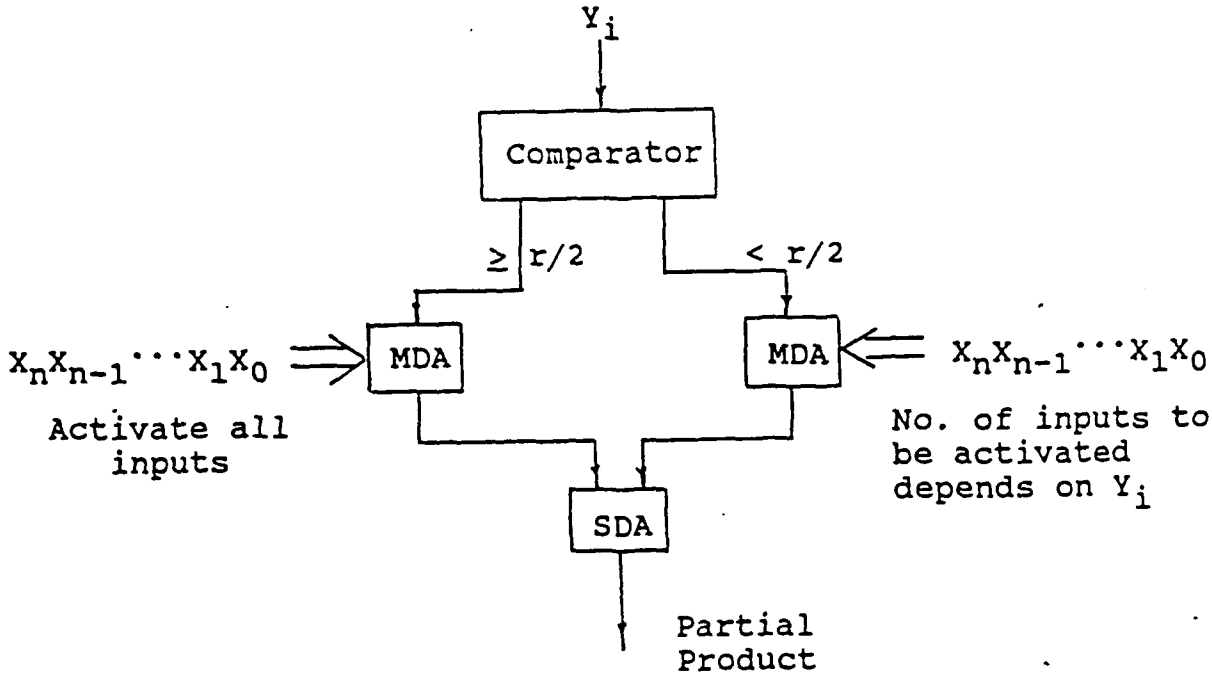


Fig. 6 Partial Product Generator Using MDA's.

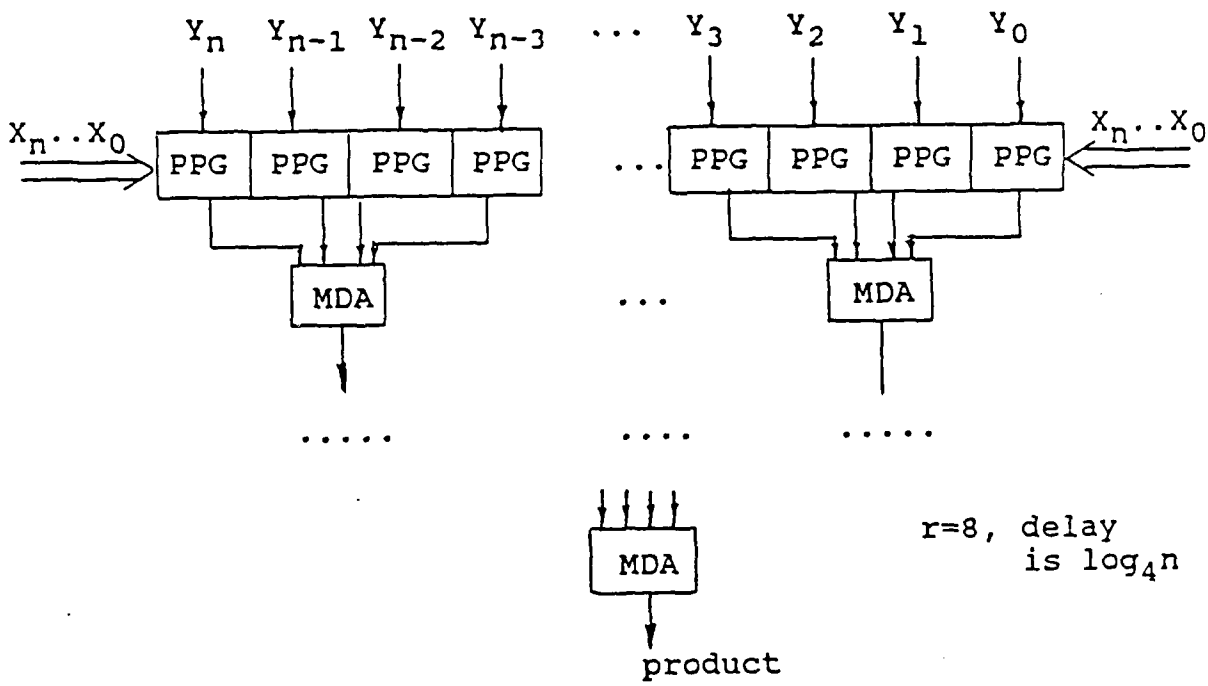


Fig. 7 Multiplier Using MDA's.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

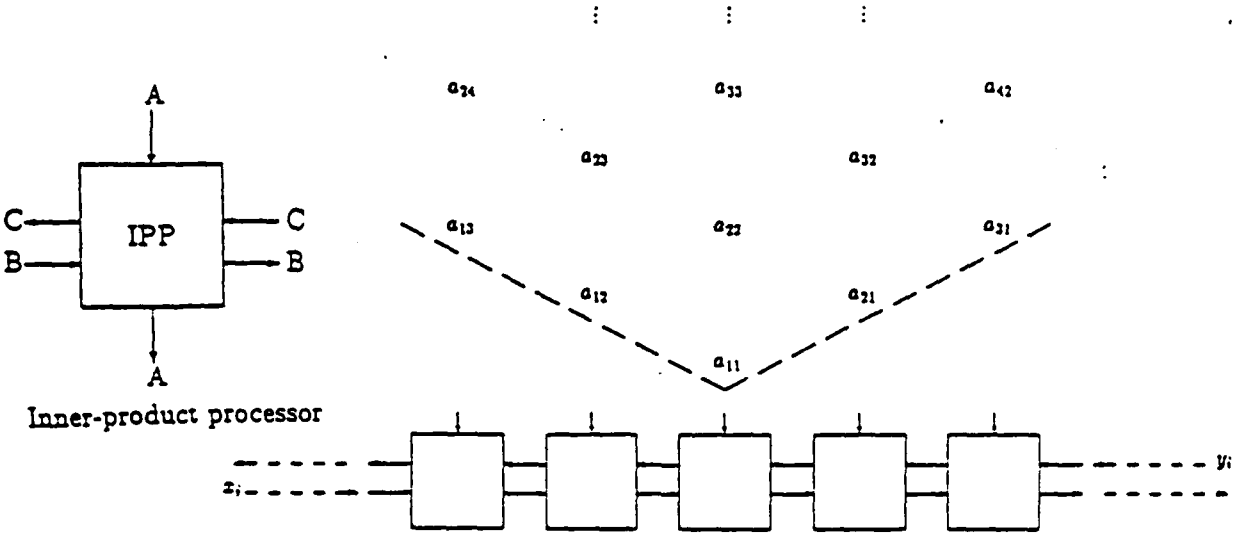


Fig. 8 Systolic Array Structure for Matrix-vector Multiplication

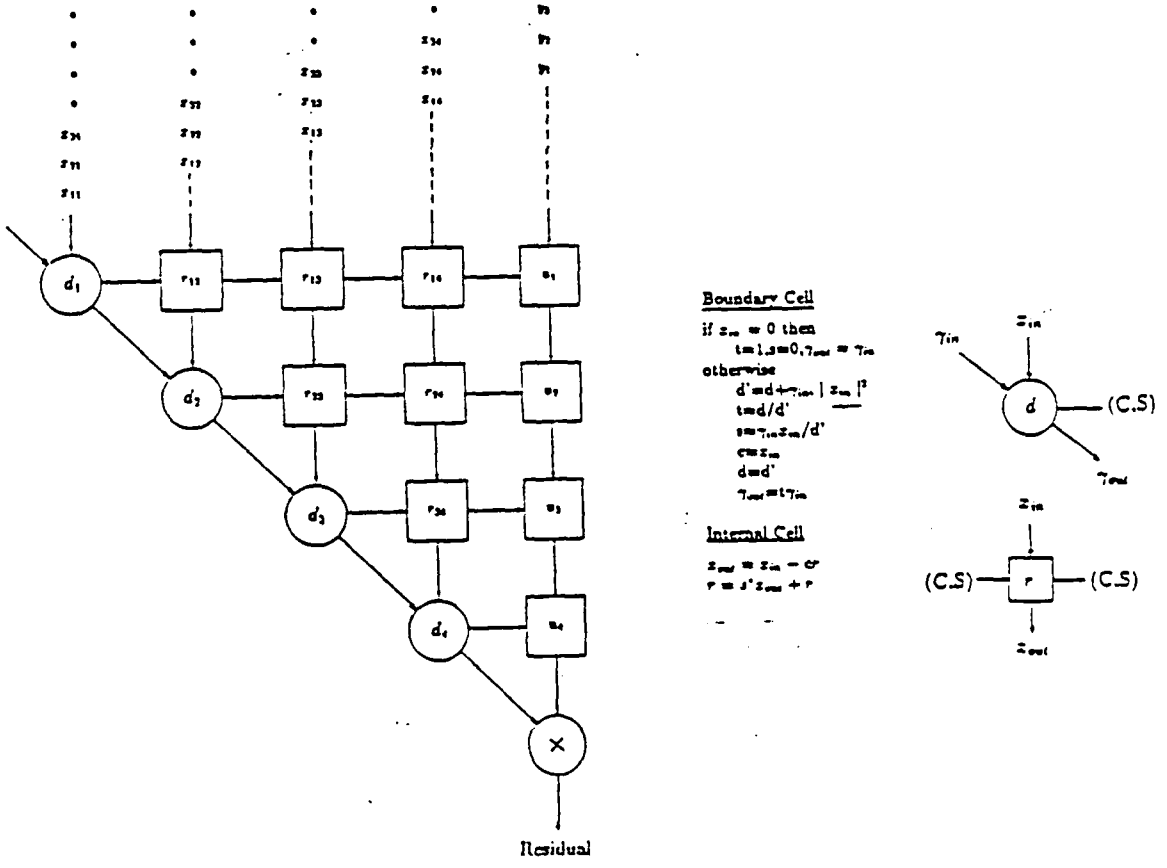


Fig. 9 Triangular Systolic Array for Adaptive Beamforming.

REFERENCES

1. S. P. Applebaum, "Adaptive Arrays," IEEE Trans. Antena Propogat., Vol. AP-24, p. 585, Sept. 1976.

2. S.P. Applebaum and D.J. Chapman, "Adaptive arrays with main beam constraints," IEEE Trans. Antennas Propogat., Vol. AP-24, p. 650, Sept. 1976.

3. R.A. Monzingo and T.W. Miller, Introduction to Adaptive Arrays, John Wiley & Sons, Inc. 1980.

4. Statement of work for Optical Adaptive Processor, PR No. c-5-2406, Rome Air Development Center, Griffiss Air Force Base, New York, May 1985.

5. A. Avizienis, "Signed-digit nuber representation for fast parallel arithmetic," IRE Trans. Electronic Computers, Ec-10,389, 1961.

6. B.L. Drake, R.P. Bocker, M.E. Lasher, R.H. Pellerson and W.J. Miceli, "Photonic computing using the modified signed-digit number representation," Optical Engineering, Vol. 25, p. 38, Jan. 1986.

7. P.A. Ramamoorthy and S. Antony, "Optical MSD adder using polarization coded symbolic substitution," Optical Engineering, Aug. 1987.

8. P.A. Ramamoorthy, S. Antony and G. Govind, "optical implementation of systolic FIR filters," Optical Engineering.

9. P.A. Ramamoorthy, S. Antony and T.A. Grogan, "Median filters - Optical implementation using symbolic substitution," SPIE's 31st Annual Intl. Tech. Symp. Optical and Optoelectronic Applied Science and Engineering, Aug. 1987.

10. N. Takagi, H. Yasuura, and S. Yajima, "High-speed VLSI multiplication algorithm with a redundant binary addition tree," in IEEE Trans, on Computers, Vol. C-34,No. 9, Sept. 1985, pp 789-796.

11. C. Mead, and L. Conway, Introduction to VLSI systems, Addison-Wesley, Reading, Mass., 1980.

12. C.R. Ward, P.J. Hargrave, and J.G. McWhirter, "A novel algorithm and architecture for adaptive digital beamforming," in IEEE Trans. Antennas Propagation, Vol. AP-34, March 1986, pp 338-346.

13. W.M. Gentleman, and H.T. Kung, "Matrix triangularization by systolic arrays," in Proc. SPIE, Vol. 298, Real-Time Signal Processing IV, 1981, pp 19-26.

14. W.M. Gentleman, "Least square computations by Givens transformations without square roots," in J. Inst. Maths Applications, Vol. 12, 1973, pp 329-336.

15. P.R. Beaudet, A.P. Goutzoulis, E.C. Malarkey, and J.C. Bradley, "Residue arithmetic techniques for optical processing of adaptive phased array radars," in Applied Optics, Vol. 25, Sept. 1986, pp 3097-3112.

FINAL REPORT NUMBER 46
REPORT NOT AVAILABLE AT THIS TIME
Dr. David Sumberg
760-7MG-113

Final Report

Implementation of Iterative Algorithms for an Optical Signal Processor

Submitted by:

Stephen T. Welstead

COLSA, Inc.

6726 Odyssey Drive

Huntsville, AL 35806

and

Department of Mathematics and Statistics

The University of Alabama in Huntsville

Huntsville, AL 35899

Date: 25 April, 1988

Period Covered: 1 April, 1987 to 31 March, 1988

Submitted to:

Rome Air Development Center, OCTS

Griffiss Air Force Base, New York 13441-5700

Attn: Dr. Vincent Vannicola

Sponsored by the

Air Force Office of Scientific Research

Conducted by the

Universal Energy Systems, Inc.

Contract No.: F49620-85-C-0013

Contents

1. Introduction

The purpose of this study is to examine the possibility of implementing an iterative algorithm such as the conjugate gradient algorithm in an optical signal processor. This research is an extension of work done as part of the Summer Faculty Research Program (SFRP) in 1986 at RADC, Griffiss AFB, NY. The period of performance covered by this report is April 1, 1987 to March 31, 1988.

The SFRP work focused on a prototype acousto-optic signal processor which was already in experimental operation as part of an RADC project (see [1,2]). This processor uses a variation of the Least Mean Square (LMS) algorithm. The goal of the current project is to investigate more powerful algorithms such as conjugate gradient that might provide improved performance for such a processor.

2. Background on the Problem

2.1 The Signal Processing Application

The particular signal processing application is adaptive noise cancellation. A main signal is received consisting of the signal of interest s(t) plus a noise signal n(t). Omni-directional side antennas receive signals $n_j(t)$, j=1,...,N. A weighted combination of delayed versions of these side signals is used to estimate the noise n(t). We denote this estimated noise by y(t). The problem is to determine the optimum combination of weights in order to minimize the difference between the estimated noise and the actual noise.

The quantity we would like to minimize is

$$E(|e(t)|^2) \tag{2.1}$$

where e(t), the so called 'error signal', is the difference between the main signal plus noise s(t) + n(t) and the estimated noise y(t), and E indicates expected value over all time with respect to some probability distribution. In practice, rather than a true expected value over all time, some finite measure or summation of recent signal history is used.

The expression (2.1) can be thought of as a functional (ie., real valued operator) of the unknown weight vector w used to form the estimated noise. It is well known [3] that the minimization of this functional is equivalent to setting its gradient equal to zero. This leads to the linear equation

$$\mathbf{A}w(x) = b(x) \tag{2.2}$$

where w(x) is the unknown weight vector function evaluated at the delay point x, b is a vector function formed from the side signals and the main signal plus noise, and A is a positive definite symmetric operator corresponding to the covariance matrix in discrete formulations of this problem (see Appendix 1 or [4] for a discussion of the derivation of the analog version of this problem).

2.2 The Least Mean Square Approach

The formulation of the quantities A and b in equation (2.2) is a formidable computational task. As a result, several approaches have been advanced which attempt to circumvent this difficulty. One of these, the least mean square (LMS) algorithm (cf., [5]), has been implemented on several optical processors ([1], [6]), including the one under consideration here. It is the performance of this algorithm that we would like to improve upon.

Although the LMS algorithm is usually thought of as an approximation of more complicated algorithms for minimizing the quantity (2.1), one can also think of it directly as an algorithm for minimizing the quantity

$$|e(t)|^2 \tag{2.3}$$

instead of minimizing the quantity (2.1). As was the case before, this minimization problem is equivalent to setting a certain gradient equal to zero. In the case of a single side signal, the gradient associated with (2.3) is proportional to

$$e(t)n_1(t-x). \tag{2.4}$$

This gradient expression has the advantage of being easy to compute. In particular, it does not involve the calculation of a covariance matrix. However, the expression (2.3) only has a minimum in the case when (2.4) is zero. This can happen only when e(t) is zero. But e(t) has the form

$$s(t) + n(t) - y(t).$$

We hope to make the quantity n(t) - y(t) zero, but in general s(t) is not zero, and so e(t) will also not be zero when there is a main signal present. This is a potential problem with LMS and we will consider it further in the next section.

Iterative processes have the general form

$$w_{i+1}(x) = w_i(x) + a_i p_i(x) \qquad (2.5)$$
$$i = 0,1,...$$

where $w_i(x)$ is the i^{th} iterative approximation of $w(x)$, $p_i(x)$ is a direction vector which indicates the direction to go in to get to the next iterate w_{i+1}, and a_i is the scalar stepsize that tells how far to go in the direction p_i.

For the LMS algorithm, we take $p_i(x)$ to be the vector given by (2.4) with $t = i\Delta t$, where Δt is the time increment between iterations. The stepsize is taken to be a sufficiently small fixed scalar a. As discussed in [4], it is possible to solve the LMS iteration process directly to obtain

$$w_k(x) = a \sum_{i=0}^{k-1} e_i n_1(i\Delta t - x), \qquad (2.6)$$

where $e_i = e(i\Delta t)$. Letting $\Delta t \to 0$, we get the analog version of (2.6), namely

$$w(x) = a \int_0^t e(s) n_1(s-x) ds. \qquad (2.7)$$

It is actually this solution, and not the iterative version of LMS, that is being implemented in the optical processors discussed in [1] and [6]. In this form, LMS is not a true iterative algorithm. Rather, it represents an approximate version of a complete solution of the minimization problem.

The advantage of LMS is the ease with which it can be implemented in a real time processor. The flow of data in such a processor is uninterrupted as the as the solution is continuously updated. This makes it particularly appealing for use in an optical processor. This is a desirable property of LMS that we should try to retain. Unfortunately, there are problems inherent in LMS that result in a degradation of performance that can reach unacceptable levels.

2.3 Problems with LMS

As mentioned in the previous section, there may be problems associated with LMS when a main signal is present (ie., signal-to-noise ratio (SNR) greater than 0). We can observe this phenomenon in the following numerical example (all numerical examples for this report were produced on a personal computer using Turbo-Pascal).

Figure 2.1 shows the performance of a numerical simulation of the LMS method in a case when the main signal s(t) is 0. The signal received at the main antenna is just a noise signal n(t) which we are attempting to cancel. In this example,

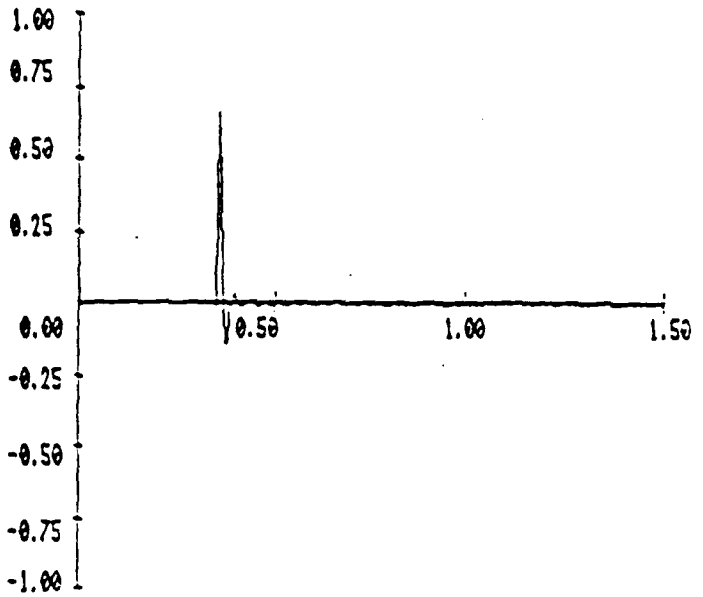


FIGURE 2.1 LMS WITH SNR = 0

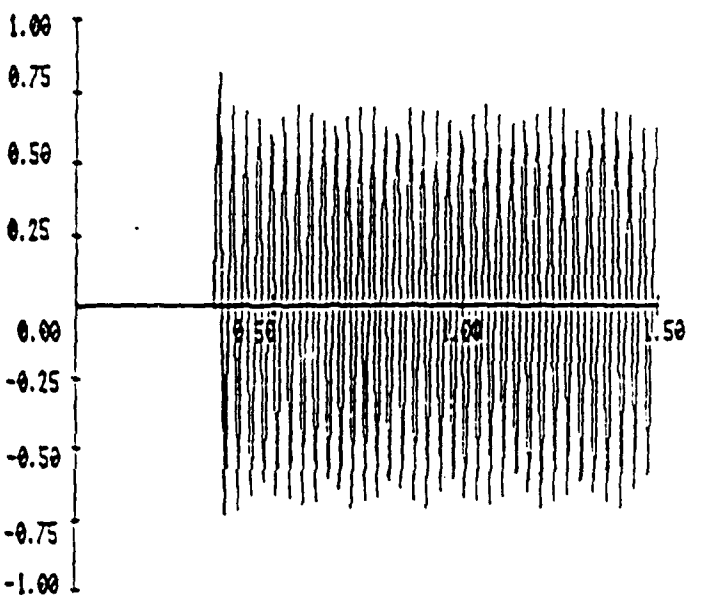


FIGURE 2.2 LMS WITH SNR = 0.5

$$n(t) = \sin(20\pi t). \tag{2.8}$$

The side antenna signal is of the form.

$$n_1(t) = \sin(20\pi t + 0.1). \tag{2.9}$$

There are 30 delay taps spread over a delay aperture of 0.3 sec. The fixed stepsize is a = 0.05. As we can see in this figure, good noise cancellation is achieved after approximately half a second.

However, when even a small main signal is present, performance deteriorates drastically. Figure 2.2 shows the effect of adding a main signal of the form

$$s(t) = 0.5 \sin(30\pi t) \tag{2.10}$$

(so that the SNR is 0.5). The graph shows

$$n(t) - y(t) \tag{2.11}$$

the difference between actual noise and estimated noise. As one can see, there is essentially no noise cancellation. This is in agreement with observed experimental results [7], citing that LMS works well in "extremely poor SNR environments". Indeed, there is no hope of it working otherwise!

Why should this be the case? Recall that

$$e(t) = d(t) - y(t)$$

where

$$d(t) = s(t) + n(t)$$

is the signal received at the main antenna. If we substitute this expression for e(t) in (2.4), and then use (2.4) as the direction vector p_i in (2.5), with $t = i\Delta t$, we obtain the following form for LMS:

$$w_{i+1}(x) = w_i(x) + a(d(i\Delta t) - y(i\Delta t))n_i(i\Delta t - x). \tag{2.12}$$

Convergence of this method implies

$$w_{i+1} \approx w_i$$

for large i, which, in turn, implies that the second term on the right side of (2.12) must converge to 0. But this implies

$$d(i\Delta t) - y(i\Delta t) \rightarrow 0$$

or, equivalently,

$$s(i\Delta t) + n(i\Delta t) - y(i\Delta t) \rightarrow 0.$$

But this quantity can never be 0 if s(t) is independent (uncorrelated) of n(t) and y(t) (which we hope is the case if we are going to avoid cancelling the main signal!). Thus, LMS is trying to annihilate a quantity that can never be zero.

To put this another way, in the case when s(t) is not identically zero, the quantity (2.3) has no minimum weight associated with it. LMS is trying to solve a problem that has no solution. The method which we introduce in the next section not only has better performance characteristics than LMS, but also completely avoids this serious drawback of LMS as a noise cancellation algorithm in the presence of a main signal.

3. Nonstationary Iterative Methods

3.1 New Approach to Iteration

We now consider a new way of incorporating iterative algorithms in a real time signal processing environment. The motivation for the approach is optical signal processing, which allows us the computational speed to consider such an approach. The uniqueness of the method lies in the fact that the data flow is allowed to drive the iterations, providing effective real time performance. Rather than perform multiple iterations on a fixed problem, which must be formulated from stored data, we allow variations in the incoming data to continuously update the problem while iterations are being performed. This is well suited to optical processing, where data storage and retrieval can be a problem, but computational speed is not. The result is an adaptive process that can significantly outperform the traditional LMS algorithm.

In contrast to the LMS algorithm, the new iterative technique deals with equation (2.2) directly, rather than an approximation of that equation. To illustrate the technique, we consider the simplest type of iterative algorithm of the form (2.5), namely the steepest descent algorithm with fixed stepsize. This algorithm has the form

$$\begin{aligned} w_{n+1} &= w_n + a\, r_n \\ r_n &= b - A\, w_n. \end{aligned} \tag{3.1}$$

The fixed scalar a is the stepsize. The sequence $\{w_n\}$ constructed from (3.1) will converge to the solution w_* of (2.2) provided

$$a < 1/M$$

where M is the largest eigenvalue of A (cf. [8]).

The usual approach in implementing an algorithm such as (3.1) is to compute A and b from the input data once, and then to regard them as fixed while the iterations are being performed. However, for our real time acousto-optic processor, it is easier to recompute A and b on every iteration, rather than to store and retrieve their values. This recomputation of A and b, however, introduces variations in their values as the iterations are being performed. Thus, it is more appropriate to write the algorithm (3.1) in the form

$$\begin{aligned} w_{n+1} &= w_n + a\, r_n \\ r_n &= b_n - A_n w_n \end{aligned} \tag{3.2}$$

where A_n and b_n are the updated versions of A and b at the n^{th} iteration.

The algorithm (3.2) is an example of a nonstationary iterative process as defined for example in [9]. In practice, one finds that A_n and b_n do in fact change on every iteration. What remains the same, however, is that the sequence of problems

$$A_n w = b_n, \quad n = 0,1,2,... \tag{3.3}$$

all have the same solution w_* for each value of n (or, at least, w_* changes slowly in time compared to the speed of the iteration process).

This makes sense in the context of our noise cancellation problem. Recall that the weight vector solution w_* represents which of the delayed versions of the side signal are to be weighted. This is not going to change from one iteration to the next. Thus, the solution does not change, even though the formulated problem changes from one iteration to the next.

When the solution does change over time, this type of process will adapt to the new solution since we are always incorporating the most recent signal data. Moreover, convergence to the new solution value should be very quick since the old solution value provides a good starting point from which the iteration process can seek the new solution.

Other iterative algorithms can also be put in nonstationary form. One improvement on the steepest descent algorithm is to optimize the stepsize at each iteration step. The nonstationary version of this algorithm has the form

$$\begin{aligned} w_{n+1} &= w_n + a_n r_n \\ r_n &= b_n - A_n w_n \\ a_n &= (r_n, r_n)/(r_n, A_n r_n). \end{aligned} \tag{3.4}$$

The nonstationary conjugate gradient algorithm takes the form

$$
\begin{aligned}
w_{n-1} &= w_n + a_n p_n \\
p_{n+1} &= r_{n+1} - c_n p_n \\
a_n &= (r_n, p_n)/(p_n, A_n p_n) \\
c_n &= (r_{n+1}, A_n p_n)/(p_n, A_n p_n) \\
r_n &= b_n - A_n w_n.
\end{aligned} \tag{3.5}
$$

Here, a_n and c_n are scalars, (,) indicates inner product, and p_n is the direction vector. In the next section, we show numerically that sequences $\{w_n\}$ generated from either (3.2), (3.4) or (3.5) will converge to the common solution w_* of the sequence of problems (3.3). In section 4.1 we look at analytical results concerning the convergence of such sequences to the desired solution w_*.

3.2 Numerical Results

This section presents the results of three numerical simulations comparing the performance of several nonstationary iterative algorithms and the traditional LMS algorithm. As mentioned previously, all numerical results were produced on a personal computer. In order to be computationally feasible on such a computer, the examples are constructed so that an exact solution is possible with a relatively small number of tap weights (we choose 6 tap weights for the iterative algorithms and 30 for LMS). In order to study the behavior and stability of the methods for larger numbers of tap weights, more computer power will be needed. For an optical processor, however, large numbers of tap weights will present no computational difficulty.

EXAMPLE 1: For the first example, we have no main signal, so that SNR = 0. The noise signal to be cancelled is

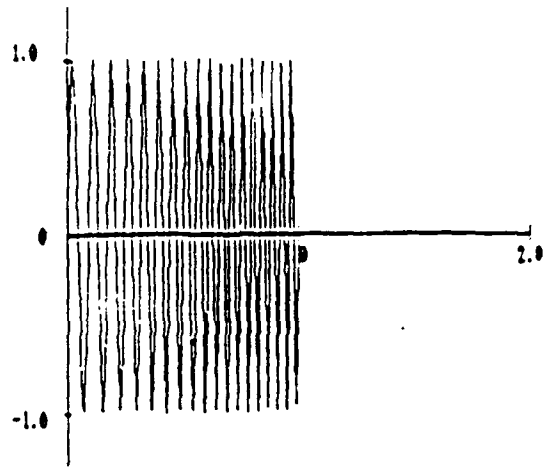
$$n(t) = \sin(20\pi t + 50t^2), \quad 0 < t < 1.$$

The graph of n(t) is shown in Figure 3.1 (a). A single side antenna receives a copy of the noise signal in the form

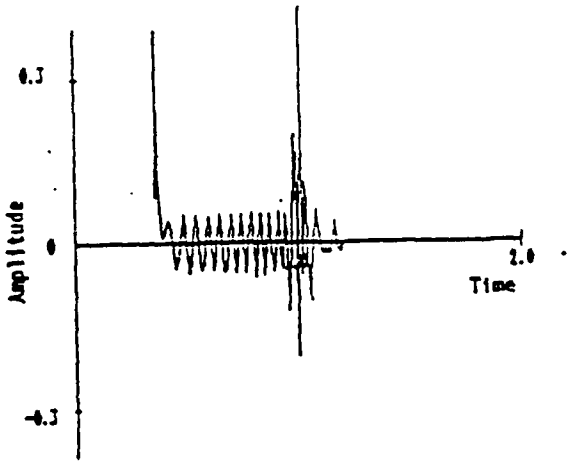
$$n_1(t) = n(t + 0.1).$$

Delayed versions of this side antenna signal are formed over a total delay aperture of R = 0.3.

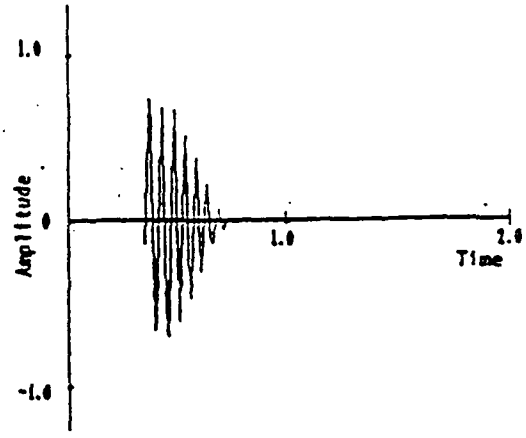
Figure 3.1 (b) shows the results for the LMS algorithm. The algorithm is run with a fixed stepsize of 0.1 and 30 delay taps. 200 iterations are used over a time interval from t = 0.35 to t = 1.3 (ie., at each iteration the current time is updated by a time increment of $\Delta t = (1.3 - 0.35)/200$. The graph



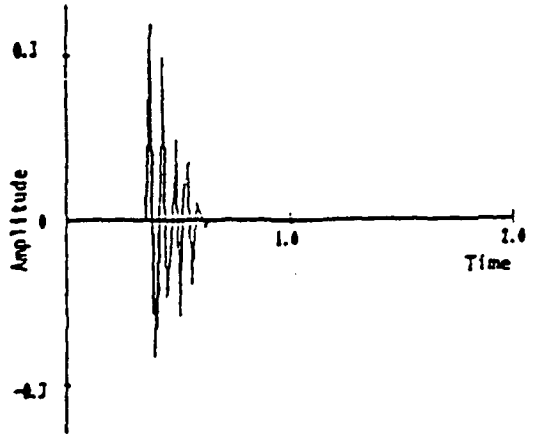
(a) NOISE SIGNAL



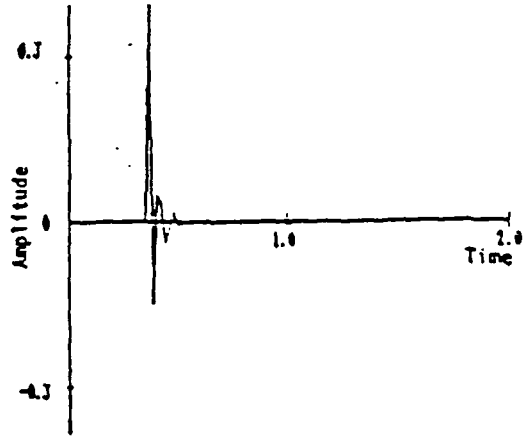
(b) LMS FILTER OUTPUT



(c) NONSTATIONARY STEEP. DESC.



(d) NONSTATIONARY OPT. STEEP. DESC.



(e) NONSTATIONARY CONJ. GRAD.

FIGURE 3.1 EXAMPLE 1

shows the difference between actual noise and estimated noise. We observe that this output noise settles down to a signal of amplitude 0.05, although the algorithm does display some problems near t=1, where the noise signal becomes compressed (higher frequency).

Figures 3.1(c)-(e) show numerical results for, respectively, the nonstationary steepest descent, with fixed and optimized stepsize, algorithm and conjugate gradient algorithm. The number of delay taps used is 6, so that the covariance matrices A_n are 6X6, and the vectors b_n have 6 components. The entries in A_n and b_n are, respectively, auto-correlation and cross-correlation functions, which are computed using integration over time. Theoretically, this integration should be performed over the time interval $-\infty$ to ∞. However, in practice this integration can only be done over a finite interval. We choose the interval from $t_0 - 3$ to t_0, where t_0 is current time. The integration is performed numerically in the simulations using a 200 point Simpson's rule.

For these nonstationary algorithms, the values of A_n and b_n are recomputed on every iteration. The simulations are run from time t = 0.35 to t = 1.3. At each iteration, the current time is updated by an amount $\Delta t = (1.3\text{-}0.35)/(\#\text{ iterations})$.

From Figures 3.1(c)-(e), one can see that in this example the nonstationary iterative algorithms provide a significant improvement in performance over the LMS algorithm. The complexity of the noise signal causes no difficulties for these algorithms. Not surprisingly, the best performance is obtained from the conjugate gradient algorithm, computationally the most complex of the algorithms.

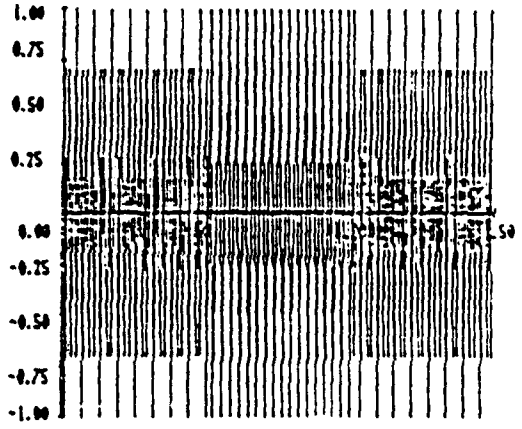
EXAMPLE 2: This is another example with SNR = 0. We consider a noise signal, shown in Figure 3.2(a), of the form

$$n(t) = \begin{cases} \sin(50\pi t) & 0 < t < 0.5 \\ \sin(100\pi t) & 0.5 < t < 1.0 \\ \sin(50\pi t) & 1.0 < t < 1.5. \end{cases}$$

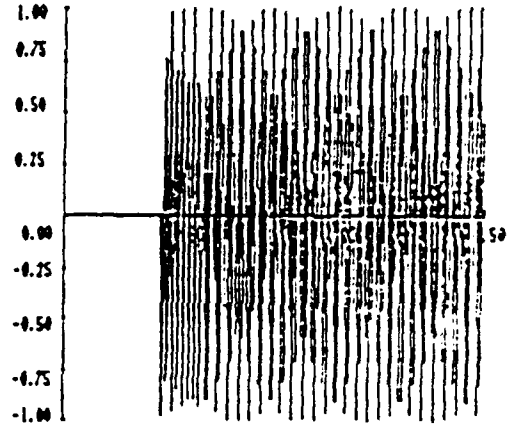
A side antenna receives a signal of the form

$$n_1(t) = n(t + 0.1).$$

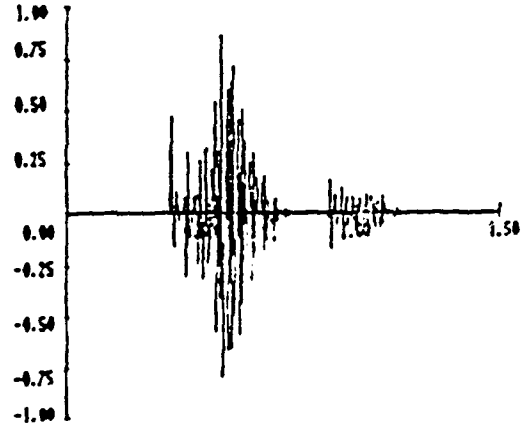
This particular noise signal was chosen to provide another example where the LMS algorithm has apparent difficulty.



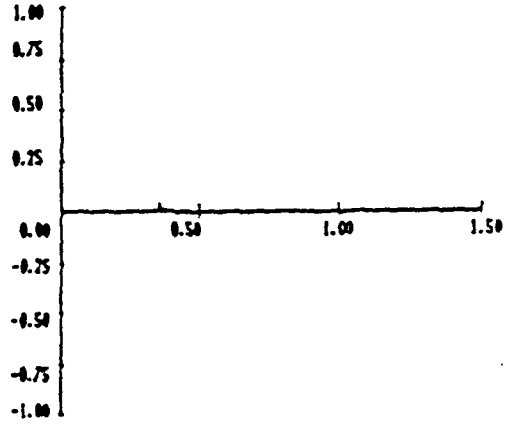
(a) NOISE SIGNAL



(b) LMS FILTER OUTPUT



(c) NONSTATIONARY STEEP. DESC.



(d) NONSTATIONARY CONJ. GRAD.

FIGURE 3.2 EXAMPLE 2

The LMS algorithm was run with a stepsize of 0.000001, with all other parameters being the same as in the previous example. Figure 3.2(b) shows the output of this algorithm. As one can see, there is essentially no noise cancellation. Larger numbers of iterations, and larger and smaller stepsizes produced no better results.

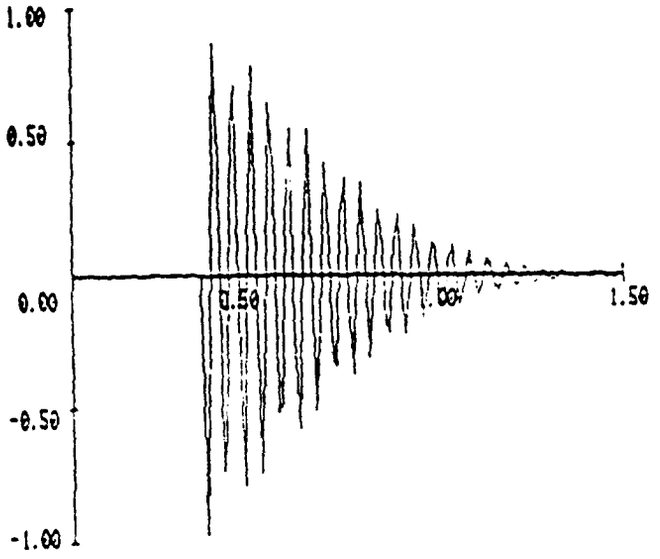
Figures 3.2(c)-(d) show the results for nonstationary steepest descent and nonstationary conjugate gradient algorithms. The noise cancellation is similar to the previous example, and is much better than LMS.

EXAMPLE 3: For our final example, we revisit the problem considered in Section 2.3. Recall that the LMS algorithm did not work at all in the presence of a main signal. Figures 3.3 (a)-(b) show the results of applying the nonstationary steepest descent with fixed stepsize and nonstationary conjugate gradient algorithms to the same problem. The noise signal is defined by (2.8), with side signal given by (2.9) and main signal given by (2.10). As one can see from the figures, the performance of these algorithms is not affected by the presence of a main signal. Figures 3.3(c)-(d) show the effect of an even larger SNR of 10. The steepest descent algorithm remains unaffected, while there is some deterioration in the performance of the conjugate gradient algorithm. It is believed that this is due to the effect of the large magnitude of s(t) on the numerical integration scheme, and not due to the conjugate gradient algorithm itself. In this example, apparently conjugate gradient is more sensitive than steepest descent to errors in the computation of A_n and b_n. This is not believed to generally be the case.

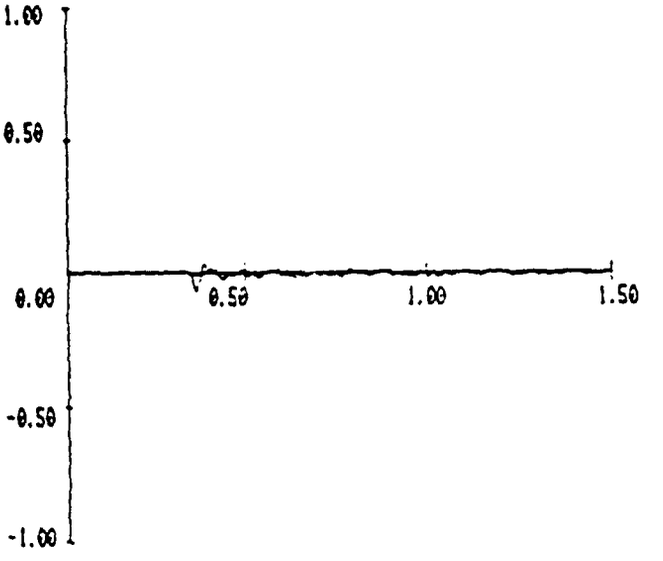
What these examples show is that there are situations where LMS does not work at all as a noise cancellation algorithm. We have shown that nonstationary iterative algorithms will work in these same situations. Since these simulations were run on a PC, the examples had to be set up so that a solution could be attained with a small number of tap weights (6). The performance of these nonstationary algorithms should be investigated on larger computers using a greater number of tap weights. Matrix pre-conditioning techniques may be necessary in this case to deal with possible ill conditioning effects.

4. Analysis

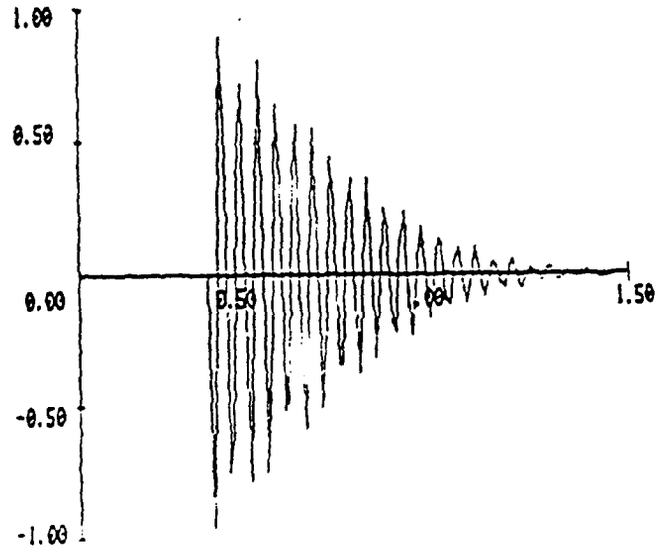
Not much is available in the literature concerning analysis results for nonstationary iterative processes of the type we are considering here. This is not surprising, since, without optical processing, such a process presents a formidable computational task. The next section contains a convergence proof for the nonstationary steepest descent algorithm. In Section 4.2, convergence results are combined with



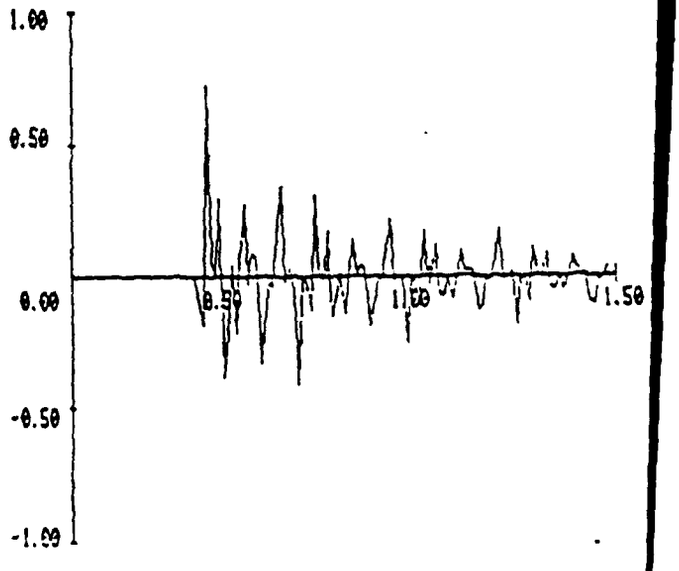
(a) NONSTATIOARY STEEP. DESC.
WITH SNR = 0.5



(b) NONSTATIONARY CONJ. GRAD.
WITH SNR = 0.5



(c) NONSTATIONARY STEEP. DESC.
WITH SNR = 10



(d) NONSTATIONARY CONJ. GRAD.
WITH SNR = 10

FIGURE 3.3 EXAMPLE 3

perturbation results to produce an error analysis for this algorithm.

4.1 Nonstationary Convergence Results

The situation we are considering is as follows. We have a sequence $\{A_k\}$ of positive definite symmetric linear operators (for example, covariance matrices) and a sequence $\{b_k\}$ of vectors such that the equations

$$A_k w = b_k, \quad k = 0,1,2,... \tag{4.1}$$

have a common solution w_*. Let the scalar a be such that

$$\| I - a A_k \| \leq \xi < 1 \tag{4.2}$$

for each $k = 0,1,2,...$, and some $\xi < 1$. The operator I is the identity operator. This is not an unreasonable condition since a similar condition is necessary for convergence of the normal steepest descent process [10]. We then have that the sequence $\{w_k\}$ generated by the process

$$w_{k+1} = w_k + a\,(b_k - A_k w_k), \quad k = 0,1,2,\ldots \tag{4.3}$$

converges in norm to w_*.

To prove this, note in the following that

$$b_k - A_k w_* = 0$$

so that we have

$$\| w_{k+1} - w_* \| = \| w_k + a(b_k - A_k w_k) - w_* \|$$

$$= \| w_k - w_* + a(b_k - A_k w_k) - a(b_k - A_k w_*) \|$$

$$= \| w_k - w_* - aA_k(w_k - w_*) \|$$

$$= \| (I - aA_k)(w_k - w_*) \|$$

$$\leq \| I - aA_k \| \, \| w_k - w_* \|$$

$$\leq \xi \| w_k - w_* \|$$

$$\leq \xi^{k+1} \| w_0 - w_* \|.$$

Since $\xi < 1$, this last term $\rightarrow 0$ as $k \rightarrow \infty$. This completes the proof.

As a corollary, we note that it suffices to replace condition (4.2) with

$$\| I - aA_k \| = \xi_k < 1 \tag{4.4}$$

for each k. We then find that

$$\| w_{k+1} - w_* \| \leq \left(\prod_{j=0}^{k} \xi_j \right) \| w_0 - w_* \|$$

and the term on the right side also $\rightarrow 0$ as $k \rightarrow \infty$ since each factor in the product is < 1. A sufficient condition for satisfying (4.4) is

$$a < m_k \tag{4.5}$$

where m_k is the smallest eigenvalue of A_k.

In Appendix 3, convergence results are given for the case when the sequence of operators $\{A_k\}$ satisfies $A_k \rightarrow A$ for some fixed operator A. However, the situation considered here, namely that the equations (4.1) have a common solution, seems to better reflect what would happen in practice. Table 4.1 shows data taken at three different time steps during one of the simulation runs discussed in the previous section. The three matrices shown here are obviously very different. What is the same is the solution $w = (0,0,1,0,0,0)$ to the three linear equations represented by these matrices and vectors.

4.2 Error Analysis

In Appendix 2 a nonstationary perturbation analysis is given for the stationary steepest descent algorithm. That is, the fixed problem

$$Aw = b$$

is solved using the usual steepest descent algorithm, and the effects of different perturbations

Covariance Matrix 1:

0.100725	-0.003987	-0.018533	-0.002036	-0.011653	0.003697
-0.003987	0.092327	-0.012011	-0.023480	-0.000029	-0.003673
-0.018533	-0.012011	0.083455	-0.019583	-0.024551	0.007741
-0.002036	-0.023480	-0.019583	0.074181	-0.024853	-0.019566
-0.011653	-0.000029	-0.024551	-0.024853	0.066359	-0.026961
0.003697	-0.003673	0.007741	-0.019566	-0.026961	0.058625

b-Vector 1:

-0.01853 -0.01201 0.08345 -0.01958 -0.02455 0.00774

Solution 1:

-0.00000 0.00000 1.00000 0.00000 0.00000 0.00000

Covariance Matrix 2:

0.133363	-0.125000	0.116637	-0.108363	0.100000	-0.091637
-0.125000	0.125000	-0.116637	0.108363	-0.100000	0.091637
0.116637	-0.116637	0.116637	-0.108363	0.100000	-0.091637
-0.108363	0.108363	-0.108363	0.108363	-0.100000	0.091637
0.100000	-0.100000	0.100000	-0.100000	0.100000	-0.091637
-0.091637	0.091637	-0.091637	0.091637	-0.091637	0.091637

b-Vector 2:

0.11664 -0.11664 0.11664 -0.10836 0.10000 -0.09164

Solution 2:

-0.00000 0.00000 1.00000 -0.00000 0.00000 -0.00000

Covariance Matrix 3:

0.167037	0.037466	-0.009107	-0.002472	-0.013037	-0.000387
0.037466	0.157761	0.036683	-0.000259	0.003787	-0.013928
-0.009107	0.036683	0.150458	0.036188	0.006093	0.010720
-0.002472	-0.000259	0.036188	0.141609	0.031116	0.008366
-0.013037	0.003787	0.006093	0.031116	0.132684	0.026026
-0.000387	-0.013928	0.010720	0.008366	0.026026	0.125200

b-Vector 3:

-0.00911 0.03668 0.15046 0.03619 0.00609 0.01072

Solution 3:

-0.00000 -0.00000 1.00000 0.00000 -0.00000 0.00000

TABLE 4.1 THREE DIFFERENT MATRIX PROBLEMS WITH SAME SOLUTION

introduced at each step of the iteration process are studied. Thus, at the n^{th} step, instead of having exactly A and b available, we assume that we are dealing with perturbed versions of these quantities:

$$A + \delta A_n$$
$$b + \delta b_n.$$

The analysis provides a bound for the difference between the perturbed iterates $\tilde{w}_n$ and the normal unperturbed iterates w_n.

In this section we apply these ideas to the nonstationary steepest descent process. As before, we consider a sequence of problems

$$A_n w = b_n, \quad n = 0,1,2,... \tag{4.1}$$

with common solution w_*. We now introduce perturbations δA_n and δb_n at each iteration step, so that we obtain a sequence of perturbed problems of the form

$$\tilde{A}_n w = \tilde{b}_n$$

where

$$\tilde{A}_n = A_n + \delta A_n$$
$$\tilde{b}_n = b_n + \delta b_n.$$

This is a particularly important problem to consider in the context of optical implementation, since we can expect errors in the formulation of A and b at each iteration step. We now determine the effect of these errors.

The nonstationary steepest descent algorithm applied to the sequence of problems (4.6) has the form

$$\tilde{w}_{n+1} = \tilde{w}_n + a(\tilde{b}_n - \tilde{A}_n \tilde{w}_n), \quad n = 0,1,2,... \tag{4.7}$$

We assume that the stepsize a has been chosen to satisfy the condition (4.2). The nonstationary process generates a sequence $\{\tilde{w}_n\}$. From a practical point of view, what we would like to know is: for large n, how far off is the perturbed iterate $\tilde{w}_n$ from the true solution w_* of the unperturbed system (4.1)?

We answer this question in several stages. First, we determine the maximum difference between the perturbed iterate $\tilde{w}_n$ and the corresponding iterate w_n from the unperturbed process (4.3). The analysis here is very similar to that given in Appendix 2, so we only sketch the details.

Define

$$\delta w_n = \tilde{w}_n - w_n, \quad n = 0,1,2,...$$

Subtracting equation (4.3) from (4.7), we find that δw_n satisfies a *nonhomogeneous* difference equation of the form

$$\delta w_{n+1} = (I - a\bar{A}_n)\delta w_n + a\, g_n, \quad n = 0,1,2,... \tag{4.8}$$

where

$$g_n = \delta b_n - \delta A_n w_n.$$

Equation (4.8) can be solved directly to obtain

$$\delta w_{n+1} = \sum_{k=0}^{n} \left\{ \prod_{j=k+1}^{n} (I - a\bar{A}_j) \right\} a g_k$$

(see Appendix 2 for the precise meaning of the noncommutative product of operators on the right). Thus,

$$\| \delta w_{n+1} \| \le a \left(\sum_{k=0}^{n} \left\{ \prod_{j=k+1}^{n} \|I - a\bar{A}_n\| \right\} \right) \left(\max_{k \le n} \|g_k\| \right). \tag{4.9}$$

Denote

$$\alpha \equiv \sup_n \| \delta A_n \|$$

$$\beta \equiv \sup_n \| \delta b_n \|$$

$$W \equiv \sup_n \| w_n \|.$$

α and β are finite by assumption and W is finite since we assume $\{w_n\}$ is a convergent sequence. Then

$$\max_{k \le n} \|g_k\| \le \beta + \alpha W.$$

Also,

$$\|I - a\tilde{A}_j\| \le \|I - aA_j\| + a\alpha$$

$$< \xi + a\alpha$$

so that

$$\sum_{k=0}^{n} \left\{ \prod_{j=k+1}^{n} \|I - aA_j\| \right\} \le \sum_{k=0}^{n} \left\{ \xi + a\alpha \right\}^k. \tag{4.10}$$

The right side of (4.10) converges to

$$\frac{1}{1 - (\xi + a\alpha)}$$

as $n \to \infty$, provided

$$\xi + a\alpha < 1.$$

Thus, in this case we have from (4.9),

$$\sup_n \| \delta w_n \| \le \frac{a}{1 - (\xi + a\alpha)} \left(\beta + \alpha W \right). \tag{4.12}$$

If we define m_j as the smallest eigenvalue of A_j and let m equal the infimum of the sequence $\{m_j\}$ then

$$\|I - aA_j\| = 1 - am_j \le 1 - am.$$

If we assume $m > 0$ and take

$$\xi = 1 - am < 1$$

the condition (4.11) becomes

$$\alpha < m$$

and the bound (4.12) can be written as

$$\sup_n \| \tilde{w}_n - w_n \| \le \frac{1}{m - \alpha} \left(\beta + \alpha W \right). \tag{4.13}$$

This implies that the perturbed process (4.7) could become unstable if the perturbations on A_n exceed m.

We can now estimate the difference between $\tilde{w}_n$ and w_*:

$$\| \tilde{w}_n - w_* \| \leq \| \tilde{w}_n - w_n \| + \| w_n - w_* \|.$$

For sufficiently large n, given $\epsilon > 0$ we can write

$$\| \tilde{w}_n - w_* \| \leq \frac{1}{m - \alpha} \left(\beta + \alpha W \right) + \epsilon. \tag{4.14}$$

This is the desired result providing the distance of the perturbed iterates $\tilde{w}_n$ from the true solution w_*. Not surprisingly, this distance depends on the size of the errors α and β, and this distance can blow up very quickly if α is close to m.

For the case of a single fixed equation, error bounds analogous to (4.14) are frequently written in terms of the condition number of a matrix. We can obtain a similar result here, if we agree to define the "condition number of the sequence $\{A_n\}$" to be the quantity

$$\Gamma \equiv \frac{M}{m}$$

where M is the supremum of the sequence $\{M_n\}$, where M_n is the maximum eigenvalue of A_n. We assume M is finite, as well as the quantity

$$B \equiv \sup_n \| b_n \|.$$

Then from (4.14) we get a bound for the relative error:

$$\| \tilde{w}_n - w_* \| / W \leq \frac{\Gamma}{1 - \left(\frac{\alpha}{m} \right)} \left(\frac{\beta}{B} + \frac{\alpha}{M} \right).$$

Thus the relative error in $\tilde{w}_n$ is proportional to the relative errors in $\{\tilde{A}_n\}$ and $\{\tilde{b}_n\}$. The constant of proportionality is dependent on Γ, the condition number of the sequence $\{A_n\}$, as well as the proximity of α to m. Once again we observe that the process can become unstable if the perturbations (α) on A_n exceed the smallest (m) of the eigenvalues of all the A_n.

5. Optical Systems

In this section we consider two approaches for possible optical implementation of the nonstationary iterative algorithms discussed in the previous sections. The first of these is a hybrid system that would use optics to do the bulk of the computational effort and an electronic microprocessor to perform the actual algorithm iteration step. This approach allows some flexibility in the choice of the algorithm, although its performance would be limited by the optics to electronics conversion. The second approach is an all optical implementation of the nonstationary steepest descent with fixed stepsize algorithm. This processor would be able to run just the one algorithm. It would, however, be an important step toward realizing all-optical implementations of the other algorithms, such as conjugate gradient, and it would provide real time performance.

5.1 Hybrid System

The first approach we consider is an electro-optic hybrid system. This system will use optics to do the hard computational task of computing the covariance matrix A_n and the vector b_n on every iteration. These computations involve correlations and integrations which can be easily accomplished optically. The iteration step of algorithms such as (3.4) and (3.5), however, involve scalar division which cannot easily be done in the optics domain. An electronic microprocessor will be used to perform this step. The use of a programmable microprocessor here will also allow the testing and comparison of different algorithms in a real signal environment. The division of tasks between optics and electronics in this hybrid processor is shown in Figure 5.1.

An overview of the hybrid system is shown in Figure 5.2. A single side signal $n_1(t)$ will pass through a tapped delay line and drive an array of light emitting diodes (LED's). The LED's are a low cost alternative to a laser system. Also, unlike lasers, the LED's have linear characteristics over a broad range in converting the input electrical signal into light, and their incoherent nature frees the system from speckle (coherent noise) present in lasers.

The LED's will illuminate an acousto-optic (AO) spatial light modulator. Figure 5.3 shows the details of the optics. The AO cell will simultaneously be driven by the same side signal $n_1(t)$, so that delayed versions of that signal will be spread across the cell aperture. This aperture should be wide enough to produce sufficient delay (about 40 μ-sec) in the side signal. This use of AO cells to produce delayed signals is similar to the techniques used in the optical signal processors of [1] and [6].

The LED's produce a vector whose components are delayed versions of the side signal $n_1(t)$. The same vector is represented in the crystal aperture of the AO cell. The result of illuminating this aperture

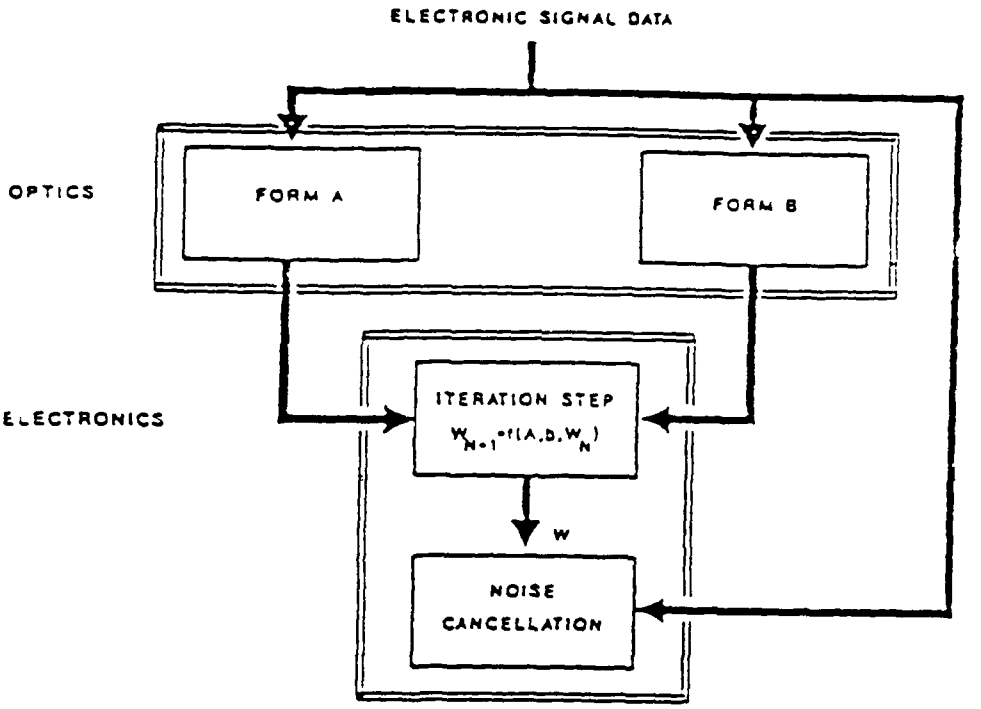


FIGURE 5.1 DIVISION OF TASKS IN HYBRID ELECTRO-OPTIC PROCESSOR

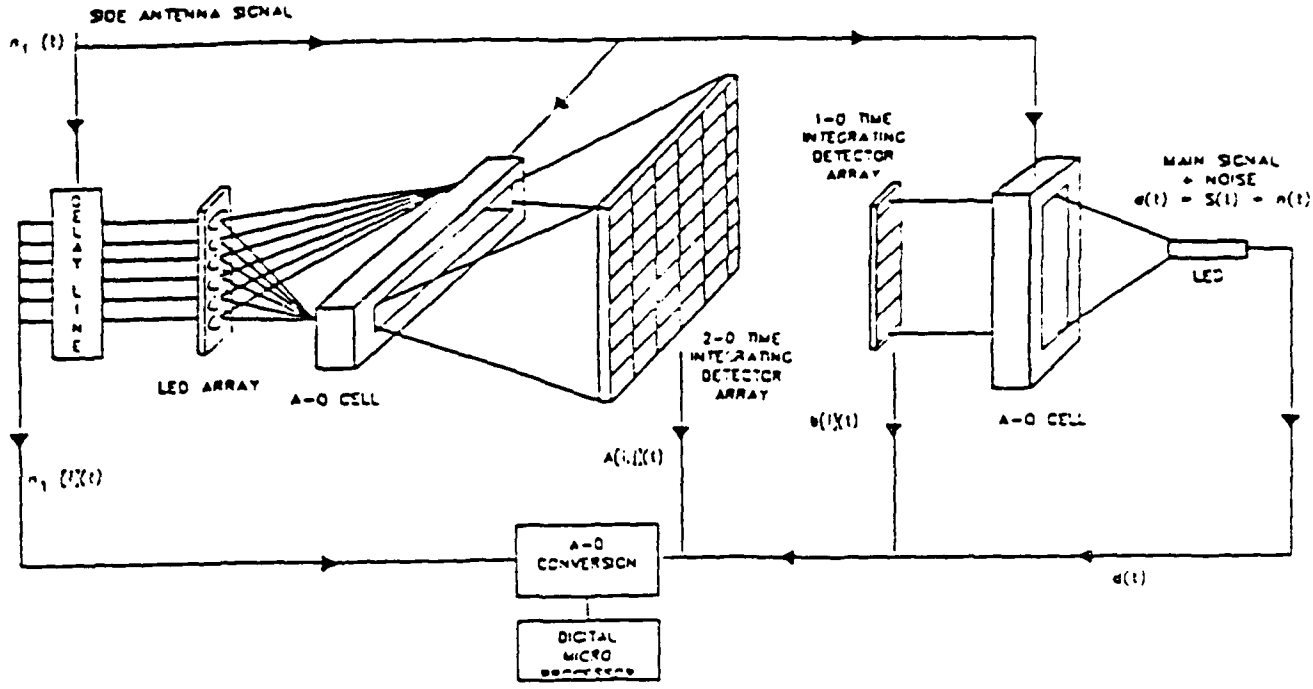


FIGURE 5.2 HYBRID SYSTEM OVERVIEW

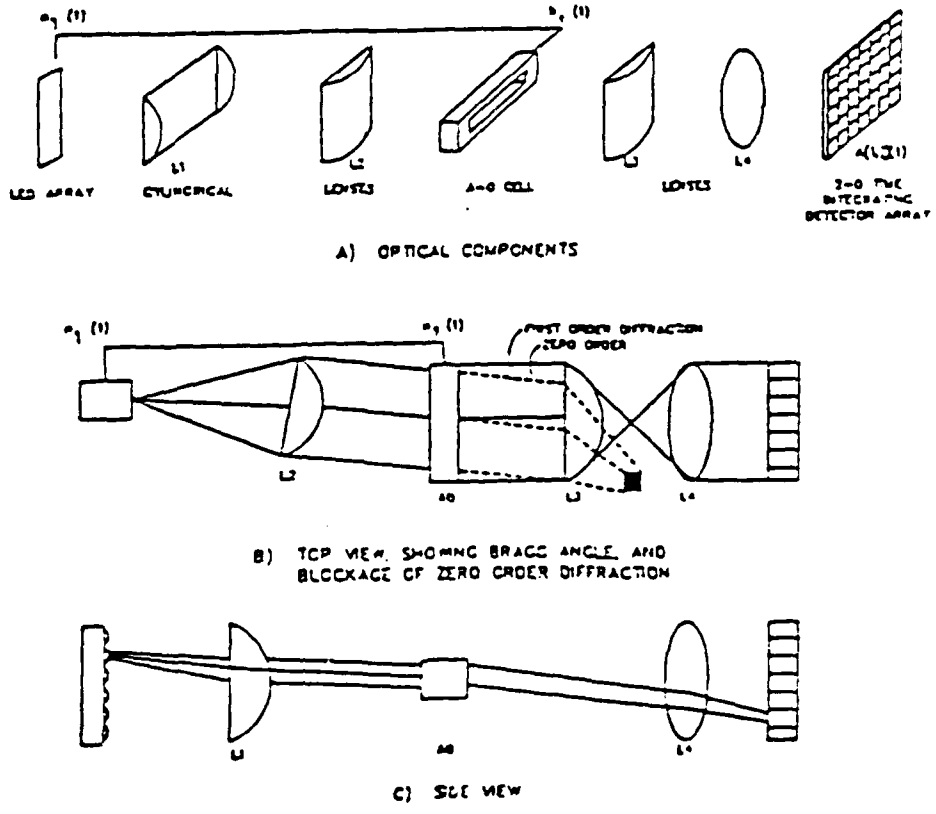


FIGURE 5.3 DETAILS OF OPTICS FOR COMPUTING A[i,j](t)

with the LED's is the outer product of these vectors, which is a matrix. This matrix is collected and time integrated by a 2-dimensional time integrating charge coupled device (CCD) detector array. The output of the detector array is the covariance matrix A_n. A frame grabber will send the matrix data to the digital microprocessor.

To construct the vector b_n at each time step, a single LED, driven by the main signal plus noise, $s(t) + n(t)$, illuminates an AO cell which is simultaneously being driven by the side signal $n_1(t)$. The resulting modulated light represents a vector whose components are the product of $s(t) + n(t)$ with delayed versions of $n_1(t)$. This light is collected onto a one dimensional CCD time integrating detector array. The output of this detector array is b_n, which is sent to the microprocessor via an analog to digital (A/D) converter board.

The main signal plus noise, $s(t) + n(t)$, as well as the delayed versions of the side signal $n_1(t)$ from the tapped delay lines, are also sent through the A/D board to the microprocessor. The iteration step and the actual noise cancellation will be performed in the digital signal domain within the microprocessor.

Such a hybrid system should be viewed as a low cost proof of principle device that could validate this class of nonstationary iterative processes for signal processing applications. The A/D conversion would limit its usefulness as a real time processor.

5.2 All-Optical System

Figure 5.4 shows a simplified system diagram for a possible optical implementation of the nonstationary steepest descent with fixed stepsize algorithm. Only one side signal is shown, although multiple side signals could be handled with a multi-channel AO cell.

AO cells are used to produce a continuum of delayed versions of the side signals, and to form products of these delayed signals with other quantities. A lens performs spatial integration. Liquid crystal light valves (LCLV) perform time integration. The weight vector w is computed in the optic domain, and the output of the system is an optical representation of the estimated noise signal y(t). This will be converted by a detector to the electronic domain where it will be recombined with the main signal plus noise, $s(t) + n(t)$, to produce the final system output, namely

$$s(t) + n(t) - y(t).$$

This signal should be close to the main signal s(t).

While the nonstationary steepest descent algorithm is not the most powerful we have considered here,

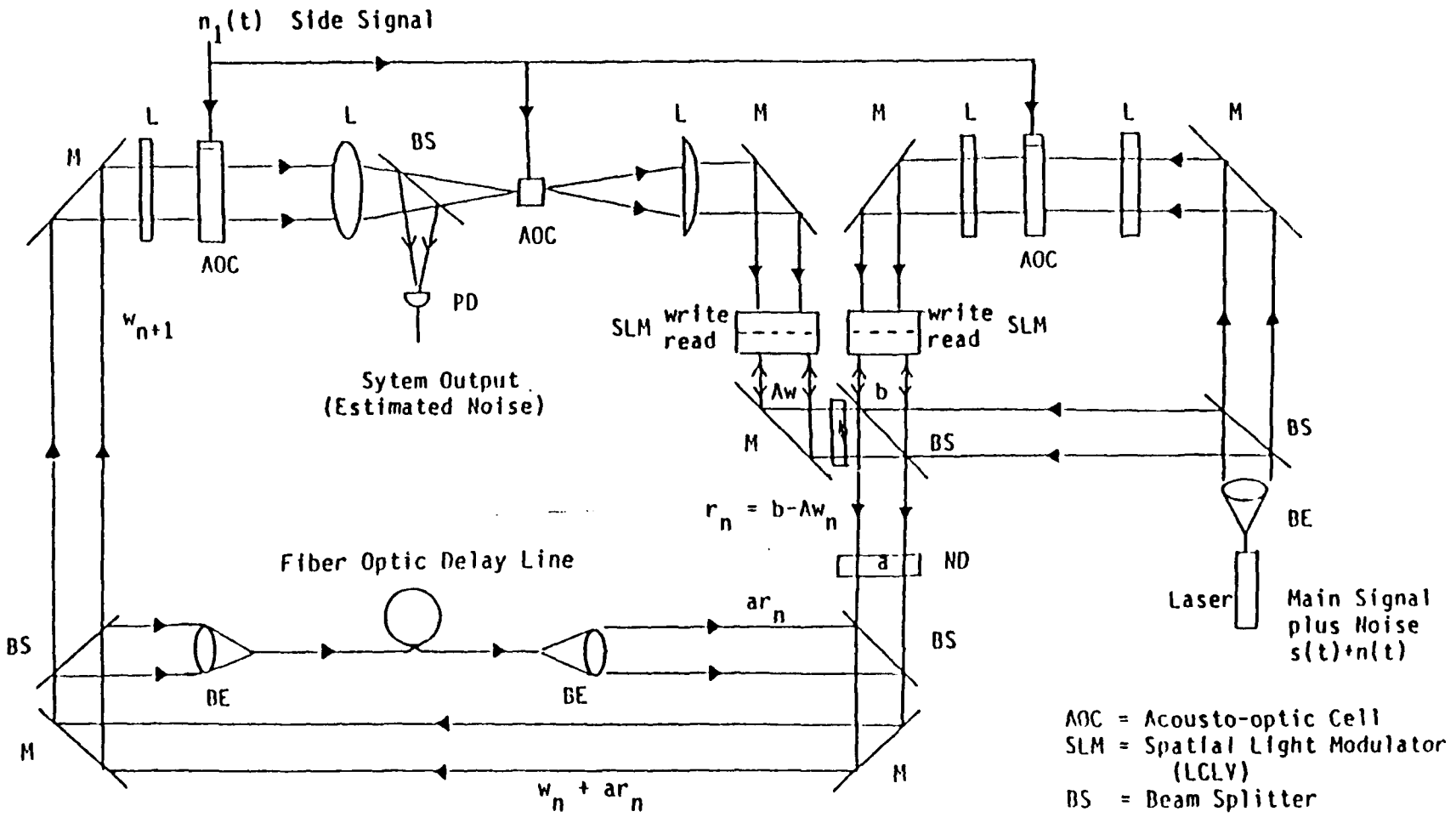


FIGURE 5.1 ALL-OPTICAL IMPLEMENTATION OF NONSTATIONARY STEEPEST DESCENT ALGORITHM

it is the simplest to implement optically. A realization of the optical implementation for this algorithm would be an important step toward opening up this whole class of nonstationary iterative algorithms to optical implementation.

6. Concluding Remarks and Recommendations

6.1 The State of the Art

Optical signal processors have already been built which can perform adaptive noise cancellation ([1] and [6]), and optical processors implementing iterative algorithms such as steepest descent and conjugate gradient have also been built, or at least proposed ([11] and [12]). So what is new about the approach that is being presented here? The optical processors that actually take in signal data and perform adaptive noise cancellation in real time are implementing some version of the LMS algorithm, and thus suffer from the performance limitations of that algorithm. The optical processors which use true iterative algorithms such as steepest descent do so on fixed matrix data, in the form of some type of mask. Thus they are not true real time signal processors, ie., they cannot formulate the matrix problem in real time and solve it. The approach we are advancing here does propose to formulate the problem in real time and solve it with the performance advantages of iterative algorithms.

6.2 Recommendations

Nonstationary iterative algorithms can provide significant advantages over LMS for adaptive noise cancellation. Optical processing will be necessary to implement these algorithms in a real time environment because of the computational load. These algorithms are good candidates for optical implementation because they take advantage of the power of optics, rather than just mimic what is already being done electronically. The technology is here now to realize optically the simplest of these algorithms, namely nonstationary steepest descent with fixed stepsize. The means to do this was outlined in the previous section. A successful optical implementation of this algorithm would open this whole class of algorithms to optics. The numerical examples of section 3 show the potential improvement possible through the use of the conjugate gradient algorithm. This algorithm could be implemented optically if a means can be found to accomplish scalar multiplication and division in the all-optic domain. The hybrid processor discussed in the previous section provides a means of validating this entire class of algorithms for signal processing applications. If improvements can be made in A/D conversion, such a processor could find practical use.

Finally, the ultimate goal of optical computing in signal processing applications should be to produce an optical processor using integrated optics, or perhaps some three dimensional analog of integrated optics (three dimensional wave guides have already been developed). Acousto-optic cells, lenses, lasers,

delay lines, and detectors have all been fabricated in integrated optics devices, with the technology for spatial light modulators lagging somewhat behind. When integrated optics technology matures we can hope to bring optical computing techniques out of the laboratory and into the field in the form of rugged, practical devices.

References

1. V.C. Vannicola and W.A. Penn, "Acousto-Optic Adaptive Processing", GOMAC Digest of Papers, Vol. X, 1984, pp. 404-409.
2. V.C. Vannicola, W.A. Penn, and M.F. Lowry, "Recent Improvements in the Acousto-Optic Adaptive Processor", GOMAC Digest of Papers, Vol. XI, 1985, pp. 477-480.
3. S.G. Mikhlin, The Problem of the Minimum of a Quadratic Functional, Holden-Day, Inc. 1965.
4. S.T. Welstead, "Preliminary Study of an Optical Implementation of the Conjugate Gradient Algorithm ", Final Report to AFOSR, 1986 USAF-UES Summer Faculty Research Program, Contract F49620-85-C-0013.
5. B. Widrow and S.D. Stearns, Adaptive Signal Processing, Prentice Hall, Inc., 1985.
6. A. Vander Lugt, "Adaptive Optical Processor", Applied Optics, Vol. 21, No. 22, 1982, pp. 4005-4011.
7. J.B. Foley and F.M. Boland, "Comparison between Steepest Descent and LMS Algorithms in Adaptive Filters", IEEE Proc.,Vol. 134, Pt. F, No. 3, 1987, pp. 283-289.
8. S.G. Mikhlin and K.L. Smolitsky, Approximate Methods for Solution of Differential and Integral Equations, Elsevier, New York, 1967.
9. J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
10. M. Hestenes, Conjugate Direction Methods in Optimization, Springer Verlag, New York, 1980.
11. A.K. Ghosh, D. Casasent, and C.P. Neuman, "Performance of Direct and Iterative Algorithms on an Optical Systolic Processor", Applied Optics, Vol. 24, No. 22, 1985, pp. 3883-3892.
12. A.K. Ghosh, "Realization of Lanczos and Conjugate Gradient Algorithms on Optical Linear Algebra Processors", SPIE Vol. 827, Real Time Signal Processing X, 1987, pp. 208-215.

Publications and Presentations

Publications supported by this contract:

S. T. Welstead, "Real Time Iterative Algorithms for Optical Signal Processing", SPIE VOL. 827, Real Time Signal Processing X, 1987, pp. 137-144.

S. T. Welstead, "Iterative Algorithms for an Optical Signal Processor", 30th Midwest Symposium on Circuits and Systems, Elsevier, 1988, pp. 490-493.

Presentations supported by this contract:

"Real Time Iterative Algorithms for Optical Signal Processing", *SPIE, San Diego, August. 1987.*

"Iterative Algorithms for an Optical Signal Processor", 30th Midwest Symposium on Circuits and Systems, Syracuse, August, 1987.

"Real Time Iterative Algorithms for Optical Signal Processing", University of Alabama in Huntsville Department of Mathematics and Statistics Colloquium, November, 1987.

Future presentation on work performed during this contract:

"Iterative Algorithms for Real Time Signal Processing", SIAM's Third Conference on Applied Linear Algebra, Madison, WI, May 23-26, 1988.

Appendices can be obtained from
Universal Energy Systems, Inc.

EXPERIMENTAL EVALUATION OF IMAGING CORRELOGRAPHY

Final Report

for the

RESEARCH INITIATION PROGRAM

sponsored by the

United States Air Force Office of Scientific Research

and conducted by

Universal Energy Systems, Inc.

4401 Dayton-Xenia Road

Dayton, Ohio 45432

Contract # S-760-7MG-109

Jerome Knopp, Principal Investigator

and

Brian K. Spielbusch, Graduate Research Assistant

March, 1989

EXPERIMENTAL EVALUATION OF IMAGING CORRELOGRAPHY

ABSTRACT

An new approach to Imaging Correlography (IC) has been found. An alternative to a more elaborate experimental system being used at the Air Force Weapons Laboratory (AFWL) was developed for low cost experimentation in IC at a university laboratory. The new approach uses optical film to record speckle patterns and an optical Fourier Transform to find the autocorrelation of the object to be imaged. The new method, the "optical approach", produces images that have resolutions comperable to those made using the AFWL system.

The new approach was used to examine the effects of glints and aberrations on IC imaging. No improvement in image resolution was found when glints were part of the object imaged. It was also found that IC imaging was unable to compensate for an aberrator mounted in the speckle recording plane.

I. INTRODUCTION

We present here results using a new type of imaging technique known as Imaging Correlography (IC). IC is a form of lenless imaging that permits image synthesis from the backscattered speckle produced by a coherently illuminated object. IC is being pursued as an alternative to large telescopes in certain imaging situations where the object to be imaged is actively illuminated by a coherent source. While on a summer research program in 1987, the authors were part were part of a research team at the Air Force Weapons Laboratory (AFWL) in Albuquerque NM that produced the first experimental verification of IC. The equipment we used at AFWL to recover an image was expensive by standards applied at the university laboratory we returned to at the end of the summer. In order to continue our experimental research at the university we developed a low cost approach to IC. The AFWL system was designed not only to prove IC but to demonstrate its implementation using fast digital image recording and processing techniques. At the AFWL laboratory we used a CCD array camera with a digital frame grabbing system. For experiments in the UMCT laboratory we were interested only in studying the effects of object glints and phase aberrations on IC; we were not interested in processing speed. Instead we used photographic film along with optical processing to replace the digital system and worked with a low power He-Ne laser. In what follows, we give a brief introduction to IC is given and a description of the simplified experimental approach used at UMCT. We shall refer to this method as the "optical approach". IC results obtained using this technique are shown and a comparison with similar

results from the AFWL setup is given. Finally, results are shown from experiments where the optical approach was used to study the effect of object glints and aberrations on IC imaging.

II. IMAGING CORRELOGRAPHY[1]

IC is a lensless imaging technique that is similar to holography. Like holography it records an interference pattern from a coherently illuminated object. This pattern is the speckle pattern one normally observes when coherent light is reflected from a diffuse object[2]. The speckle represents the coherent interference of many point scatterers on the surface of the object. Unlike conventional holography where a reference wave is used to form the hologram[3], the object to be imaged is in a sense it's own reference. Because the object and hence the reference wave are both unknown, special processing techniques are used to recover the image.

Image synthesis using IC makes use of an interesting relationship first described by Goldfisher[4]. He showed that the autocorrelation function of the illuminated object's brightness distribution can be obtained from the average power spectrum of a laser speckle pattern. (The term "brightness distribution" means the object's irradiance distribution had the object been illuminated with an incoherent light source.) Since the inverse Fourier transform of the autocorrelation of the object's brightness function is equal to the squared-modulus of the Fourier transform of the brightness function[5], an image of the object can be obtained if the phase associated with this Fourier transform can be determined. To obtain this phase we use the Fourier modulus estimated from the speckle data together with a iterative transform algorithm of the type previously demonstrated by Fienup[6] [7]. Once the phase associated with the Fourier modulus is determined, the image is recovered by inverse transforming the synthesized Fourier plane data. Because the image is recovered from an estimate of the

power spectrum of the object's brightness distribution, we find that the recovered image is unspeckled, even though the data for IC was obtained from measurements of speckle intensity.

In the ideal model for IC, it is assumed that the object is optically rough, so that its microscale surface height is random and comparable in size with the wavelength of light. In this case, the reflected laser light is randomly (and coherently) dephased, and the speckle in the recording plane is fully developed. In practical situations this assumption is often violated due to polarization changes that primarily appear to effect the speckle contrast. In the discussions that follow we shall assume ideal case.

Estimating the Autocorrelation

To obtain a perfect autocorrelation, all the speckle in an infinite plane produced by an object must be recorded and transformed. In practical circumstances only a finite portion of the speckle can be recorded and an estimate for the autocorrelation be obtained. The larger the speckle pattern processed, the better estimate for the autocorrelation will be. In many typical situations the space bandwidth product is low and the information from a single speckle record is insufficient to provide the resolution desired in the recovered image. For example, at AFWL a CCC array television camera was used to collect speckle data and we found that a single frame was inadequate. An alternative, in these cases, is to collect many independent frames of speckle data. Each frame of data (i.e. realization) can produce a rough estimate of the autocorrelation. To reduce the roughness of this estimate, the average of the estimates is

found. This average, in the limit for a large number of frames, approaches the exact autocorrelation. In recording each frame of data, the camera must view separate independent nonoverlapping areas in the speckle plane. The amount of space required between different regions of the recording plane is related primarily to the surface features of the object. The roughness means any view change in the object presents a different set of point sources. It also ultimately determines resolution which is also related to the scale of the roughness.

Formally, the nth realization of the observed speckle intensity $I_n(u)$ may be expressed as the squared modulus of the Fourier transform of the complex object field:

$$I_n(u) = |F_n(u)|^2 = |\mathfrak{F}\{f_n(x)\}|^2, \qquad (1)$$

where $\mathfrak{F}\{\ \}$ denotes a Fourier transform operator.The field reflected by the object is $f_n(x) = |f_o(x)|\exp[j\theta_n(x)]$, $|f_o(x)|$ is the object's field amplitude reflectivity (Assuming uniform object illumination.), and $\theta_n(x)$ is the phase of the nth realization of the reflected object field associated with the object's surface height profile. In the above expression, x represents a two-dimensional spatial (or angular) coordinate vector in object space; u represents a two-dimensional coordinate in the measurement plane. An estimate of the autocovariance of the measured speckle pattern may be computed as follows from N realizations of the laser speckle intensity:

$$C_I(\Delta u,N) = \mathfrak{F}\{\ N^{-1}\sum_{n=1}^{N}|\mathfrak{F}^{-1}\{P(u)\ [\ I_n(u) - I\ ]\}|^2\} \qquad (2)$$

Where I is the average intensity of the observed speckle pattern, Δu is a vector separation in the measurement plane, and P(u) is a binary function denoting the region R of the measurement plane over which the

speckle pattern is observed and is defined as follows:

$$\begin{aligned} P(u) &= 1, \quad \text{for } u \in R \\ '' &= 0, \quad \text{elsewhere.} \end{aligned} \tag{3}$$

In the limit as N, the number of independent observed speckle patterns, approaches infinity, one can use the moment factoring theorem for circular-complex Gaussian fields to show that

$$|\Gamma(\Delta u)|^2 = \lim_{n\to\infty} N^{-1} \sum_{n=1}^{N} [\, I_n\,(u + \Delta u) - I\,]\,[\, I_n(u) - I] \tag{4}$$

where $\Gamma\,(\Delta u) = \mathfrak{F}\,\{\,|f_o(x)|^2\}$ is the Fourier transform of the object brightness distribution (i.e., $\Gamma(\Delta u)$ is the mutual coherence function of the speckle field in the measurement aperture, evaluated at field points separated by a vector (Δu).) The ability to invoke the circular-complex Gaussian moment theorem above follows from the fact that the observed speckle field is circular-complex Gaussian, since the speckle pattern is fully developed. In the limit $N \to \infty$, the estimated autocovariance of the speckle intensity observed over the measurement aperture P(u) is given by

$$\begin{aligned} C_I(\Delta u) &= \lim_{N\to\infty} C_I(\Delta u, N) \\ '' &= \mathrm{OTF}(\Delta u)\ |\ \Gamma\ (\Delta u)\ |^2 \end{aligned} \tag{5}$$

where OTF is the optical transfer function, and can be related to P; i.e. $\mathrm{OTF}(\Delta u) = P(u)\ \tilde{P}(u)$, where $\tilde{}$ denotes an autocorrelation. This result demonstrates that $C_I(\Delta u)$ provides an estimate for $|\Gamma(\Delta u)|^2$, the power spectrum of the object brightness function. The square root of $|\Gamma(\Delta 2)|^2$ is an estimate of the Fourier modulus of the object's brightness distribution.

Retrieving the Phase[6]

Once the modulus of the Fourier transform of the object's brightness function is obtained, the phase must also be found to reconstruct the object. Described here are phase retrieval methods that work for general objects with additive noise. If we consider an arbitrary object field f(x) then its Fourier transform is given by :

$$\begin{aligned} F(s) &= |F(s)|\exp|j\psi(s)| = \mathfrak{F}(f(x)) \\ &\equiv \iint_{-\infty}^{\infty} f(x)\exp(j2\pi u\cdot s)dx, \end{aligned} \tag{6}$$

where the vector position x represents a two-dimensional spatial coordinate and s the two dimensional spatial frequency. For typical objects, f(x) is a real, nonnegative function. The problem is to find an object that is consistent with the nonnegativity constraint as well as the constraints imposed by the estimate of the Fourier modulus. These constraints can be used to find the phase using a retrieval technique. Among the most popular techniques for phase retrieval are iterative approaches. We will briefly describe the algorithms we used in IC image recovery.

The Error Reduction Approach

Among the simplest algorithms for phase retrieval is the error-reduction approach. At the kth iteration, $g_k(x)$, an estimate of the object, is Fourier transformed; the Modulus of this transform is set equal to the estimated modulus and the result is inverse-Fourier transformed, giving the image $g_k'(x)$. Then the iteration is completed by forming a new estimate of the object that conforms to the object-domain constraints as follows:

$$\begin{aligned} g_{k+1}(x) &= g_k'(x), \quad g_k'(x) \geq 0 \\ '' &= 0, \quad g_k'(x) < 0 \end{aligned} \tag{7}$$

One additional constraint that may be enforced is a limitation on the object size based on the size of the autocorrelation. This is done by restricting the object to zero outside a closed boundary. This additional constraint often increases significantly the convergence rate and in some case the uniqueness needed for a solution. We will describe such a constraint, the triple intersect method, later. The iteration process can be started by using a complex sequence of random numbers for $g_1(x)$.

The Input-Output Approach

In an attempt to speed up the convergence, the more powerful input-output approach was developed by Fienup. This method differs from the error reduction approach only in the object-domain operation where a feedback is used. Instead of modifying the last output, we can modify the previous input to form the new input using feedback in those regions where the output is nonnegative. The feedback strength is determined by the feedback factor β given in Eq. (7). Note the feedback is directly proportional to the strength of the negative value.

$$\begin{aligned} g_{k+1}(x) &= g'_k(x), && g'_k(x) \geq 0 \\ '' &= g'_k(x) - \beta\, g'_k(x), && g'_k(x) < 0 \end{aligned} \qquad (8)$$

The Combined Approach

It has been found based on a considerable amount of experience that alternating between the error reduction method and the input-output method every few iterations has proven to work better simply using one method for all the iterations. This might seem to suggest that the methods are compensatory. While feedback speeds convergence it may also introduce some instability, the trick is the right

approach at the right time. We will give the precise details of the procedure we used in the combined approach later.

III. THE OPTICAL APPROACH TO IC

To produce a good estimate for the autocorrelation, the number of speckle realizations may be large or a single recording of a large area of speckle can used. The choice is dictated by the recording device. A camera like the one used in the AFWL experiments has a SBWP on the order of 5×10^4. By comparison a typical frame of 35 mm. film used in black and white photography has a SBWP of about 10^6. This means that a low cost photographic film has a SBWP that permits recording a significant amount of speckle information. In our approach to IC we decided to use film instead of a television camera. There are three good reasons for this. Using film it was possible to:

(1) collect all the speckle information needed in a single recording; this avoided the problems of rotating the target to obtain many independent speckle patterns as was done in previous AFWL experiments.

(2) have long recording times and use a lower power laser; in the AFWL experiments the television frame rates required much higher illumination levels be provided by an expensive Argon laser.

(3) find our autocorrelation directly using optical processing to find the Fourier transform of the speckle intensity; this avoided storing large numbers of speckle frames and avoided a large amount of digital processing.

The last item is significant since storing a single frame of speckle data from the television camera involved over 250 kbytes of storage and an array processor to carry out the Fourier transforms (One for

each frame of data.). We used an optical Fourier transform, this meant the autocorrelation was obtained after the film record was optically transformed in a single step. We also found we were able to use a negative instead of a positive transparency without any problems. The use of a negative effects primarily the mean or DC value in the Fourier transform. The effect is observed as a spike in the center of the autocorrelation and is removed when the autocorrelation is processed. (This spike also occurs in the all digital approach used at AFWL.) Once the autocorrelation was optically obtained, it was frame grabbed and stored on a computer to be used in a phase retrieval algorithm to recover the image. The frame grabbing operation is performed only once in this case using a conventional television camera.

IV. EXPERIMENTAL ARRANGEMENTS

The experimental arrangements used to find the autocorrelation with the optical approach were simple. Only a conventional 35 mm. camera and a He-Ne laser with a spatial filter are needed. This simplicity is possible because we were not constrained to working with a system designed for low light levels and rapid framing rates. The situation in the our laboratory allowed long speckle recording times, since case we were only interested in evaluating certain glint and aberration effects independently of photon limited detection. In addition to the optical approach we also discuss arrangements needed to compare the optical and digital approaches.

Optical Estimation of the Autocorrelation

Figure 1 shows the method used to record the speckle, photograph the autocorrelation function of the object and record the image of the autocorrelation. The object to be imaged was illuminated with a He-Ne laser; the speckle field from the object was then recorded on a 35mm. camera without a lens (See Fig. 1(a)). Kodak TMAX-100 film was used record the speckle. Care was taken not to exceed it's resolution. The size of the target, and the distance to the camera had to be chosen to properly scale the highest fringe frequency in the speckle pattern. If a target has a largest dimension l and is located a distance z from film plane, then the highest spatial frequency associated with the field can be determined by the interference pattern between the objects most widely separated points, i.e.the distance l. This implies a highest fringe frequency of l/z in the field. Since the film records the intensity of the scattered speckle and is proportional to

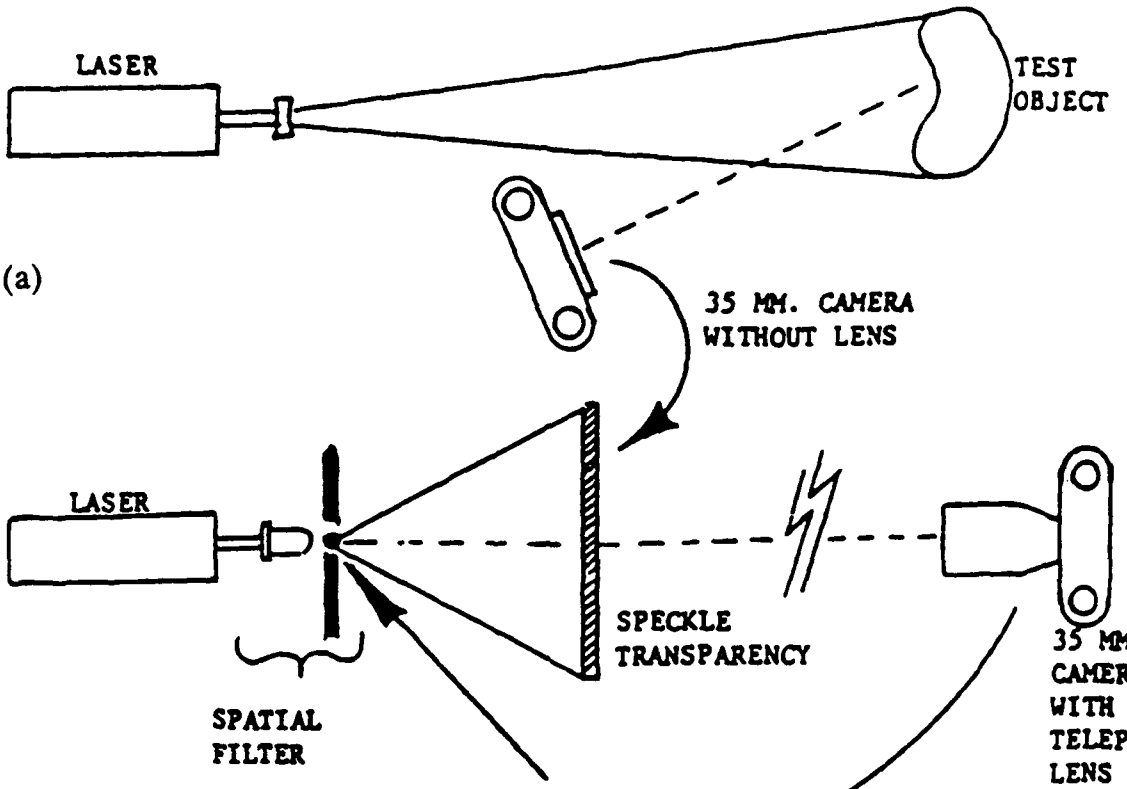


Figure 1(a)

Figure 1(b)

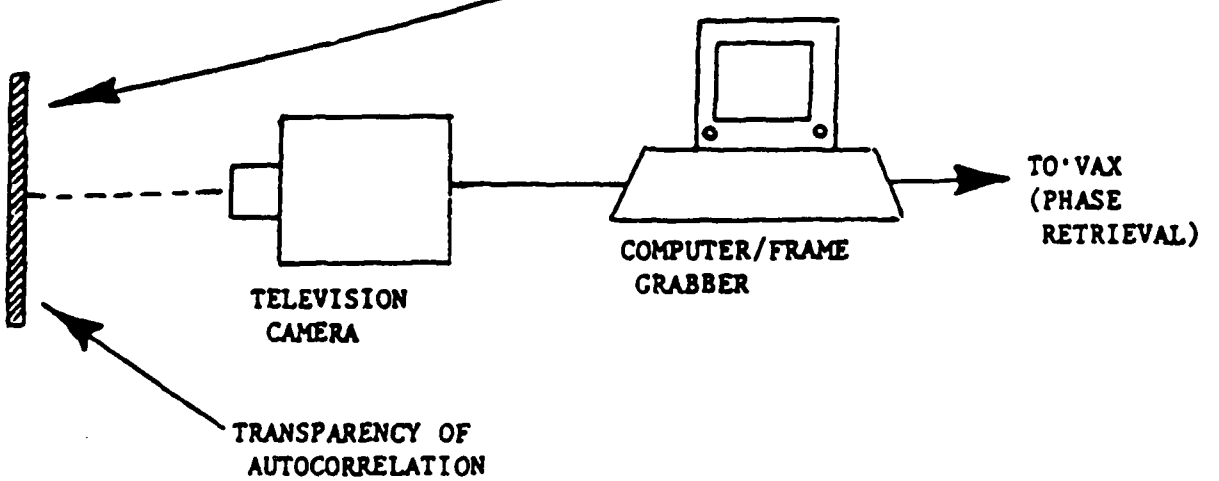


Figure 1(c)

Figure 1. The optical approach to I.C. (a) Recording the speckle, (b) Fourier transforming to find the autocorrelation (c) Digitizing the autocorrelation

magnitude of the field squared the fringe frequency recorded on film is doubled. Therefore, the film resolution must be $2l / \lambda z$, or greater. We determined the film limitations experimentally by observing the brightness of the autocorrelation. This autocorrelation was observed using the optical Fourier transform arrangement shown in Fig. 1(b). A He-Ne laser, with a spatial filter, was used to produce a point source. This point source was then viewed while looking through the speckle negative with a 300 mm.telephoto zoom lens. A virtual Fourier transform was then observed in the plane of the point source[11]. This transform, the autocorrelation of the object, was then photographed.

In using this arrangement we found it necessary to use a range of test exposures to properly adjust the film gamma. We also found the choice of film significant. We had first tried Kodak Plus X pan film and found it lacked the contrast and hence the linerity with respect to the intensity that was needed. When negatives were made with Plus X we were able to observe higher order harmonics in the autocorrelation. We found that switching to Kodak TMAX-100 reduced the nonlinearities to an insignificant level. We found exposure times on the order of 2 minutes were needed for the 3 milliwatt laser used in illuminating the test object.

The test object was made from a beaded reflector material made by sold under the tradename of Scotchlite. Scotchlite was used by us in previous AFWL experiments in IC to solve two problems: low target reflectivity and depolarization of the laser light. It has a surface coated with a large number (more than $10^5 / in^2$) of retro-reflecting microspheres with an effective diameter averaging about 50 microns.

(This was estimated from the Airy-like farfield pattern it produces.) The reflectivity was about of 12% with most of the object light reflected within a 10 degree cone. The retro-reflecting beads do not significantly effect the polarization of the target return. We were not interested in adding the polarization issue to our experiments, we preferred as ideal an object as possible.

In making objects, the backing on the Scotchlite was used to mount it on a glass background. The specular reflection off the glass could be directed away from the camera; this meant the object scene had an extremely low noise background. A negative lens was used to expand the beam from the laser just enough to fully illuminate the Scotchlite object. For all our results the object shown in Fig. 2 was used. Its greatest dimension is approximately 1.5 cm. in the vertical

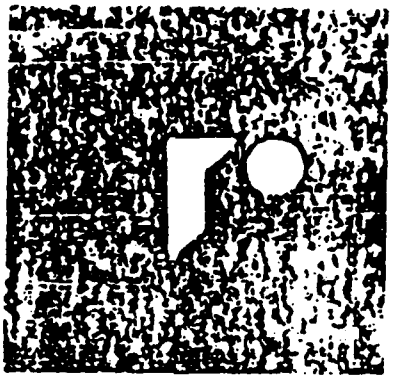


Fig. 2. Object to be imaged using IC

direction. The speckle was usually recorded at a distance about 1.5 meters from the target. This represented the closest distance we could get to the target and suggests the highest useful spatial frequency recorded was about 33 lines/mm. The TMAX film was developed in Kodak HC110 developer using the standard developing procedures recommended by Kodak. The exposure was adjusted experimentally by visual

inspection of the autocorrelation. A positive of the autocorrelation was made on Kodak variable contrast paper and imaged with a video camera, digitized and stored in the computer for image retrieval as shown in Fig. 1(c). The autocorrelation is shown in Fig. 3. The step of making a positive as opposed to directly putting the autocorrelation into the video camera was recommended to us based on previous experiences in digitizing imagery at the University of Texas at El Paso[12]. This procedure was found to give the best net gamma.

Fig.3 The autocorrelation.

Comparison of Optical and Digital Approaches

In order to make comparisons with conventional digital IC, speckle from the target was recorded using a General Electric CID 512 video camera that was available at the AFWL. The camera contained a square CCD array that was 7.68 mm on each side and was made up of 512 by 512 pixels. A single frame of TMAX film had a SBWP that was about times 12 times that of the CCD array, therefore we collected 12 speckle patterns with the television camera to estimate the autocorrelation using the digital approach. The IC image made from these speckle frames was used for comparison with the IC image made using the optical approach. In order to collect 12 speckle patterns

the object was slightly rotated horizontally to move the speckle field or the camera (which was sitting on an optical jack stand) was raised or lowered to move vertically through the speckle field. The object at AFWL was illuminated with an Argon laser operating on its .488 micron line at about 3 watts. We recorded 12 independent speckle patterns using 4 different angular positions of the target and 3 different heights on the camera jack stand. Each 512 by 512 speckle speckle pattern was windowed with a 256 by 256 window (The speckle patterns were windowed, to eliminate cross correlation in side by side speckle patterns). The mean was then subtracted off and digitally Fourier transformed to produce a coherent (rough) autocorrelation. All of the 12 autocorrelations were averaged together to obtain an estimate of the autocorrelation of the object.

Collection of Camera Speckle from the Photographic Negative

We also decided to try another type of image comparison. The speckle data from the negative used in the optical approach was imaged directly into the CCD array and digitally recorded. We were curious to discover if the image recovered using the negative speckle would be as good as that recovered using other approaches. If it was, it would indicate a wide range of tolerance for nonlinearity. The negative speckle data did not have the proper gamma for digital intensity processing; it was collected for field processing using a lens. In the case of the digital approach we wanted a gamma near 1, while in the optical approach we desired a gamma near 2. We expected a good deal of nonlinearity if we processed it directly as intensity data. Besides examining the effects of non linearities, using the same speckle data

in the camera and the optical processing technique could be helpful in examining camera artifacts. We will bring up this issue when we examine our experimental results later. We also wanted to keep a digital record of the negative data for future experiments; these experiments might include examining the effects of changing the gamma using computer simulations.

A slide of the speckle negative was placed into a X-Y translation mount and the imaging arrangement was scaled so that 12 separate areas of the 35mm. negative could each be individually imaged onto the CCD array. The total of all 12 of the areas imaged included the entire slide image. Each frame of negative speckle was grabbed by a computer and stored. After all 12 patterns were stored, the same processing procedures were applied to the negative speckle that had previously been applied to the positive speckle data that was collected directly by the camera. Each frame was windowed, its mean subtracted, then a fast Fourier transform was applied. The results were averaged to obtain an estimate of the autocorrelation.

V. DIGITAL PROCESSING PROCEDURES

In this chapter we describe the details for the digital processing procedures that were used. The actual algorithms used at AFWL are constantly updated by researchers at both AFWL and at the Environmental Research Institute of Michigan (ERIM). The procedures described start with a coarse 32 x 32 array in the first stages of the iteration algorithm and then the array resolution is increased. This approach increases the convergence speed by reducing the time involved in the Fast Fourier transform processing. The algorithms also incorporate a significant amount of practical experience handling noise. This includes subtraction techniques to eliminate camera background noise as well as Weiner filtering. These algorithms represent the state of the art reached within the last couple of years.

Digital Processing of the Autocorrelation

Step 1

Low value video camera noise was removed by subtracting off 1% of the peak value from the whole array and then setting negative values to zero.

Step 2

The object is real and nonnegative, therefore, the power spectrum and autocorrelation should be real and nonnegative. This nonnegativity constraint was applied 5 times each to the power spectrum and autocorrelation after each transformation. This application of this constraint will be discussed more later with respect to our experimental results. The modified autocorrelation was

than fast Fourier transformed one more time to obtain an estimate of the power spectrum, of the object. The square root of this spectrum was used as an estimate of the Fourier modulus.

Step 3

The power spectrum was multiplied by a Wiener filter:

$$W(u) = \frac{\Gamma(u)^2}{\Gamma(u)^2 + En} \qquad (11)$$

Where $\Gamma(u)$ is the power spectrum, and En is the power spectrum of the noise and approximated by a constant. The noise constant was chosen to be 20% of the peak value of the power spectrum. The noise value was determined by visual inspection of the Fourier modulus.

Step 4

The triple intersect method shown in Fig. 4 was applied to the autocorrelation to obtain a starting guess for the object boundary in the phase retrieval process. The points inside the boundary were initially given random values. In applying the triple intersect method the autocorrelation is first thresholded based on visual inspection of the autocorrelation (Fig. 4(A)). Then a cut is positioned visually to intersect two border points (Fig. 8(B)), two more cuts are then made going through the border points to the centroid of the thresholded autocorrelation (see Fig. 4(C)),. The triangular area contained within the three intersecting lines is saved and border information from the discarded area is then added to the edges of the triangle produced by the center cuts (Fig. 4(D)). Pixels may also be added around this new shape (i.e. the object is coated) to obtain a desired size. The size is usually chosen to be roughly half

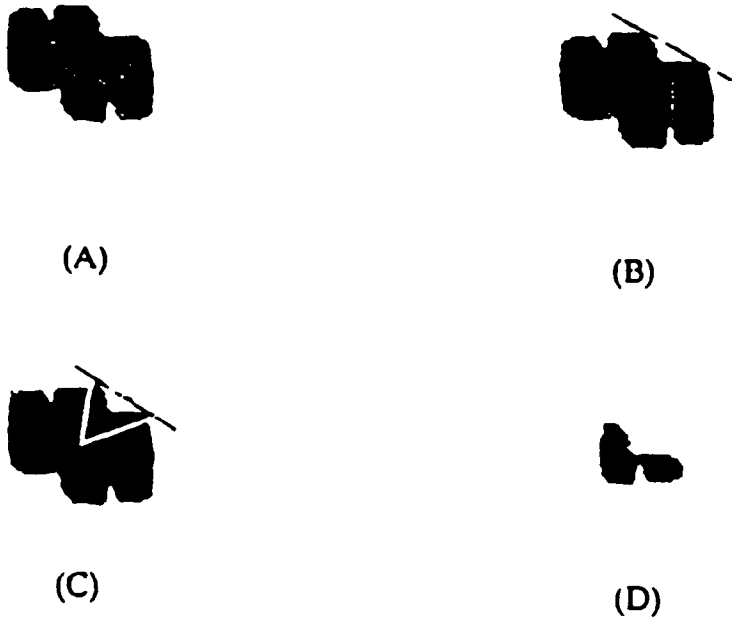


Fig. 4. Triple intersect method. (A) Thresholded autocorrelation, (B) intersection of the two border points, (C) the three intersecting lines, (D) resulting initial guess.

the size of the autocorrelation. This can be done by eye or it can be done be adjusting the radial averages.

Iteration Procedures

Starting with only the center 32 by 32 of the Fourier modulus, multiplied by a triangular window, those zero's come to the edges. Then the following iterations were applied:

10 cycles of 1 input-output (β=0.7) and 1 error reduction.

1 cycle of 20 input-output (β=0.7) and 4 error reduction, with the triangular window opened 1 pixel after each iteration. Then the array size was then increased to 64 by 64 and the iterations were continued as follows:

2 cycles of 100 input-output (β=0.7) and 5 error reduction.

1 cycle of 20 input-output (β=0.5) and 4 error reduction, with the triangular window again opened 1 pixel after each iteration. Then the iteration procedure was changed to:

2 cycles of 100 input-output (β=0.5) and 5 error reduction.

The array size was finally increased to 128 by 128 and 1 error reduction iteration was done, to double the size of the object. (Continued iterations with 128 by 128 arrays not needed because of low signal to noise beyond 64 by 64)

VI. RESULTS AND ANALYSIS

Speckle data was recorded and processed to obtain IC images using the three different methods described previously. The methods are compared here with each other and with "truth" data. The truth data was made by using a digitally simulated object that was similar to the real object used in the experimental data. The computer object was made by first imaging the object used with the CCC array camera and taking care to properly scale it to approximately the size of the expected speckle image. (The scaling was done by comparison with images recovered from the experimental speckle data.) Using Fast Fourier Transform processing on the computer the digitized image was used to get a "truth autocorrelation" for the object (Fig. 5(A)) and a "truth Fourier modulus" (Fig. 5(B)). The truth Fourier modulus was used in the same phase retrieval procedure used with the laboratory data. The truth data image is shown in Fig. 5(D).

Comparison of Results Using the Different Correlography Methods

Figure 5 contains all the key data used to evaluate the different IC methods. The images shown are black/white reversed. By this we mean the darkest areas in the image correspond to the highest intensity levels. This is especially convenient for plotting on a laser plotter using a halftone scheme to show shades of gray. The first column of figures contains the truth data as previously discussed. However Fig. 5(c) is the starting guess used with all the images. The second, third and fourth columns contain the results for the negative speckle, conventional correlography and the optical approach respectively. In these three columns the top row shows the autocorrelation, the second

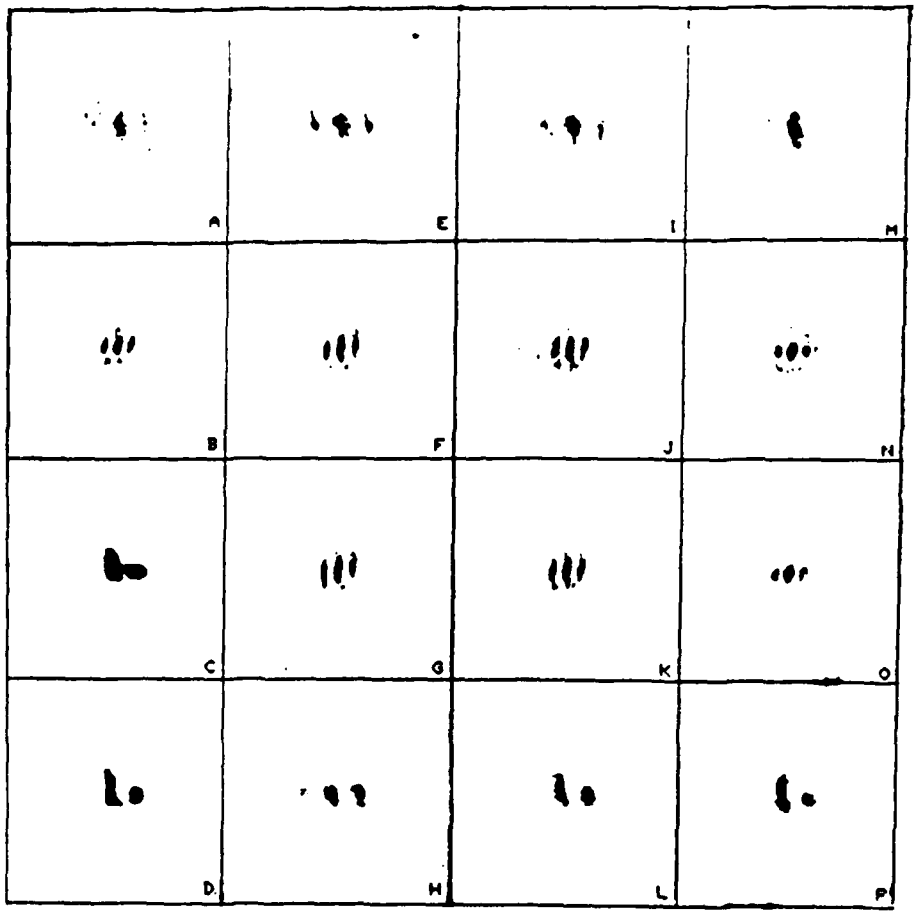


Fig. 5. Results using the different correlography methods. (A) Truth autocorrelation. (B) Truth Fourier modulus. (C) Mask used for guess, to start phase retrieval. (D) Reconstruction, with the truth Fourier modulus. (E) Noncoherent autocorrelation using negative speckle method. (F) Fourier modulus, before Wiener filtering, negative speckle method. (G) Fourier modulus, after Wiener filtering, negative speckle method. (H) Reconstruction, negative speckle method. (I) Noncoherent autocorrelation, conventional correlography. (J) Fourier modulus, before Wiener filtering, conventional correlography. (K) Fourier modulus, after Wiener filtering, conventional correlography. (L) Reconstruction, conventional correlography. (M) Noncoherent autocorrelation, optical method. (N) Fourier modulus, before Wiener filtering, optical method. (O) Fourier modulus, after Wiener filtering, optical method. (P) Reconstruction , optical method.

row shows the Fourier modulus, the third row shows the modulus after Wiener filtering and the fourth row shows the recovered images. Note the distinct difference in the autocorrelation obtained using the optical method. It appears to lack strong secondary peaks on either side of the main lobe. This shows up in a weak fringe structure in its Fourier transform modulus. This suggests a large amount of low pass filtering is probably associated with the film MTF. However its recovered image is quite competitive with the image obtained by the conventional approach. The implication of this is that the significant image information is contained in the low frequencies. Despite the presence of high frequency lobes in the conventional image the edges are not any sharper. It is interesting that the negative speckle information also shows a similar Fourier modulus. None of the Fourier modulii compare particularly well with the truth set. Note also the faint secondary orders in both the negative and conventional autocorrelations. This also shows up faintly as twin image effects in the final images. This suggests nonlinearities in the data. As previously mentioned this was to be expected with the negative data but since it also shows up in the conventional data might suggest nonlinearities in the camera. If there are nonlinearities in the film it could have been suppressed by the lowpass filtering. The final image for the negative is as expected; it is poor. Although it might be possible to adjust the effective gamma of negative data by rescaling it on the computer to improve the image, we have not yet tried this interesting exercise. The images for the conventional and the optical approach are comparable and show similar resolutions. Both of these images also show resolutions that appear to be within a

factor of three of the resolution that can be expected using a conventional imaging system with the same aperture, resolution and image size.

VII. EXPERIMENTS AND RESULTS WITH GLINTS AND ABERRATORS

We present here the results of experiments in IC using targets with glints present and targets that were imaged through an aberrating media.

The effects of glints on targets is of interest since a glint adds a reference source that in effect produces a Fourier transform hologram of the object. This implies and that Fourier transforming the speckle pattern can reproduce the object. However, this is only true if the glint is widely separated from the target. If the glint is embedded in the target the Fouier transform images are overlapped with the autocorrelation. Even with the overlap, the presence of strong image structure intuitively suggests the possibility of faster convergence and improved resolution. In the glint experiments discussed here, we will examine these issues.

It is well known that a thin phase aberrator next to the plane of a hologram does not effect the image reconstructed from the hologram. This is because the object wave and the reference wave pass through the same phase aberration and the same phase shift is added to both the object wave and the reference wave. Since hologram fringe structure is dependent only on the phase differential between the object and the reference waves at every point, it is unchanged when the same phase shift is added to both of them. This begs the question in the case of IC concerning its potential in eliminating phase aberrations near the recording plane. One may wish to argue that the situation in IC is identical to the situation in holography where the object is in effect its own reference. Furthermore, on even simpler argument can made. If a thin phase aberrator is present during the

recording process it cannot possibly change the field magnitude but only the phase at a given point and hence any intensity recording must be unaffected. It would appear to some that this experiment is not really of interest, since the outcome should be obvious. As we shall see, reality is often surprising.

Glint Experiments

Using the optical approach and the same target that was used previously we tried three experiments with glints added to the targets as shown in Figs. 6,7 and 8. In Fig. 6 a single glint was added near the target while in Fig. 7 the single glint is moved further away to obtain better separation of the holographic image in the autocorrelation. In Fig. 8, two glints are added. The glints consisted of upholstery tacks with spherical chrome heads. The head radii were small with focal lengths on the order of an inch. In working with the glints we found it necessary to vary the distances slightly between the camera recording the speckle and the object to balance as much as possible the strong returns from the glint and the Scotchlite material. These adjustments are reflected in scale differences in the final images shown in the results.

Glint Results

The IC imaging results using glints are shown in Fig. 9. The first column shows the truth data with no glints for comparison. The second column shows results for one near glint, the third column the results for one far glint and the last column the results for two glints. All results are based on 12 frame processing using the same starting guess previously used (i.e. the triple intersect object.).

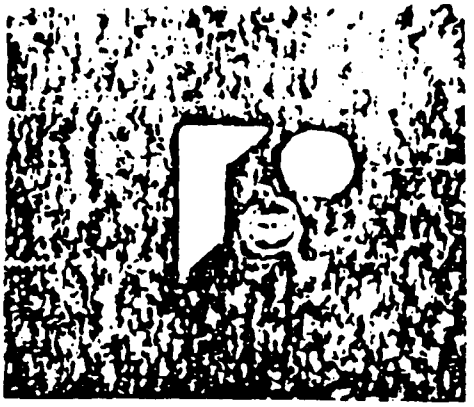


Fig. 6. Object with one glint near.

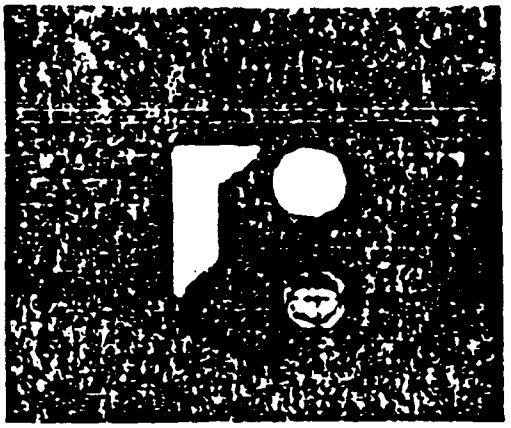


Fig. 7. Object with one far glint.

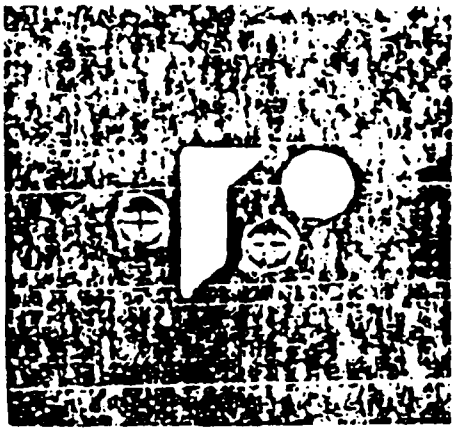


Figure 8. Object with two glints.

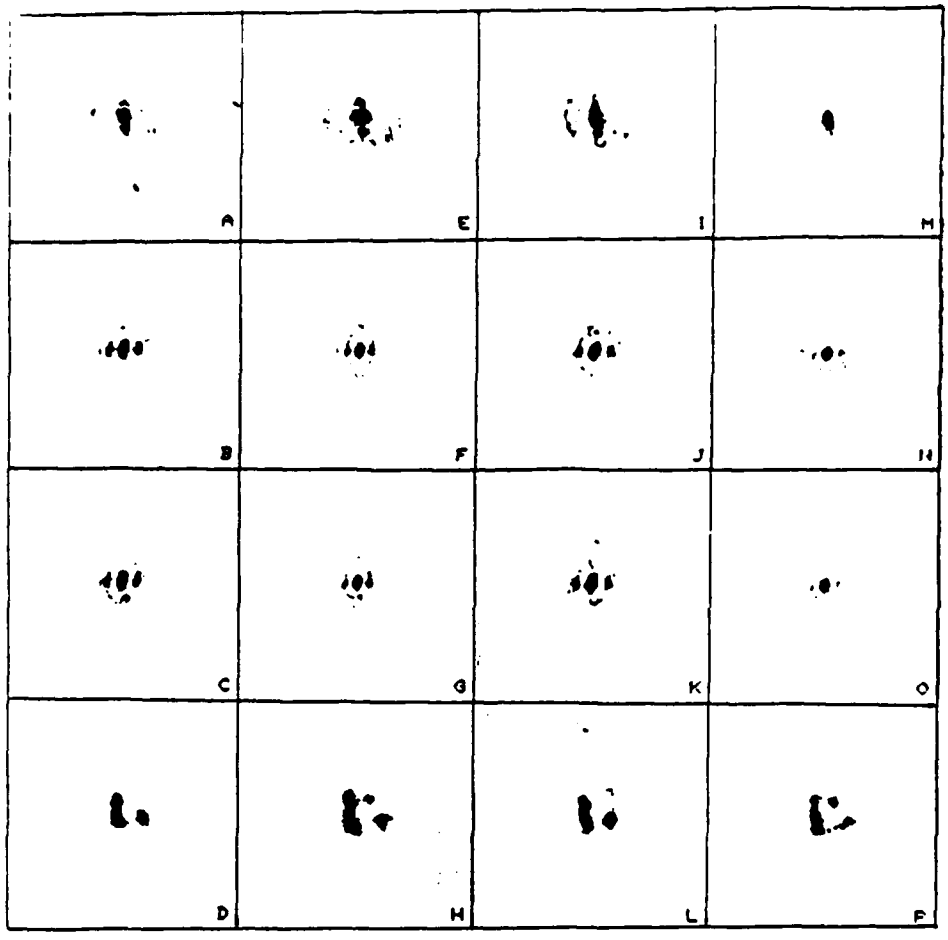


Fig. 9 Results with glints, (A) Noncoherent autocorrelation, no glints, (B) Fourier modulus, before Wiener filtering, no glints, (C) Fourier modulus, after Wiener filtering, no glints, (D) Reconstruction, no glints, (E) Noncoherent autocorrelation, one glint near, (F) Fourier modulus, before Wiener filtering, one glint near, (G) Fourier modulus, after Wiener filtering, one glint near, (H) Reconstruction, one glint near, (I) Noncoherent autocorrelation, one far glint, (J) Fourier modulus, before Wiener filtering, one far glint, (K) Fourier modulus, after Wiener filtering, one far glint, (L) Reconstruction, one far glint, (M) Noncoherent autocorrelation, two glints, (N) Fourier modulus, before Wiener filtering, two glints, (O) Fourier modulus, after Wiener filtering, two glints, (P) Reconstruction, two glints

The first row shows the autocorrelations, the second row and third rows the filtered and unfiltered Fourier modulii and the fourth row the recovered images.

The recovered images do not really show any striking improvement in resolution. In fact although the near glint results shows the glint clearly in its reconstruction, the far glint is not visible in its image and one glint appears missing in the two glint case. The Scotchlite parts of the object are apparent but the resolution does not appear to be enhanced as was hoped for. There is no clear case for an image improvement in these data. Furthermore the erratic recovery of glints in the images suggests the possibility that the specular nature of the glints may be significant.

Another striking difference in the results is the effect of the two glints on the autocorrelation and the Fourier modulus. This result may be a dynamic range problem in which the strong glint returns produce dominant low frequency terms that suppress the side lobes in the autocorrelation. If this is the case, it would suggest that glints can create additional problems in recovering the image. Although it is difficult to argue that there are significant differences in the recovered images. Most observers to which the authors have shown the data feel the two glint image looks worse or even slightly distorted.

Epoxy Aberrator Experiment

In order to investigate the effects of an aberration on IC image, a thin aberrator was constructed that could be mounted in the film plane of the speckle recording camera. The aberrator was made by coating a thin piece of clear acetate about .005 inches thick with a transparent layer of epoxy. The coating had an average thickness about .005 inches but was smeared, using a finger, to produce a phase aberrations that were quite severe, at least several thousand waves. When the aberrator was held against a planar object such as text on a sheet of paper its presence was not discernable unless the aberrator was moved approximately 1/8 inch from the paper plane. A photograph of the test object made through the aberrator (See Fig. 10) shows the severity of the aberration. The effect of the aberrator on speckle intensity was expected to be negligible provided it was in the film plane. The aberrator was mounted directly in the camera film plane and was in intimate contact with the 35 mm. film. We then recorded speckle data through the aberrator.

Figure 10. Object photographed through the aberrator

Aberrator Results

We found our first surprise when we looked at the Fourier transform of the aberrator autocorrelation data. We found it severely

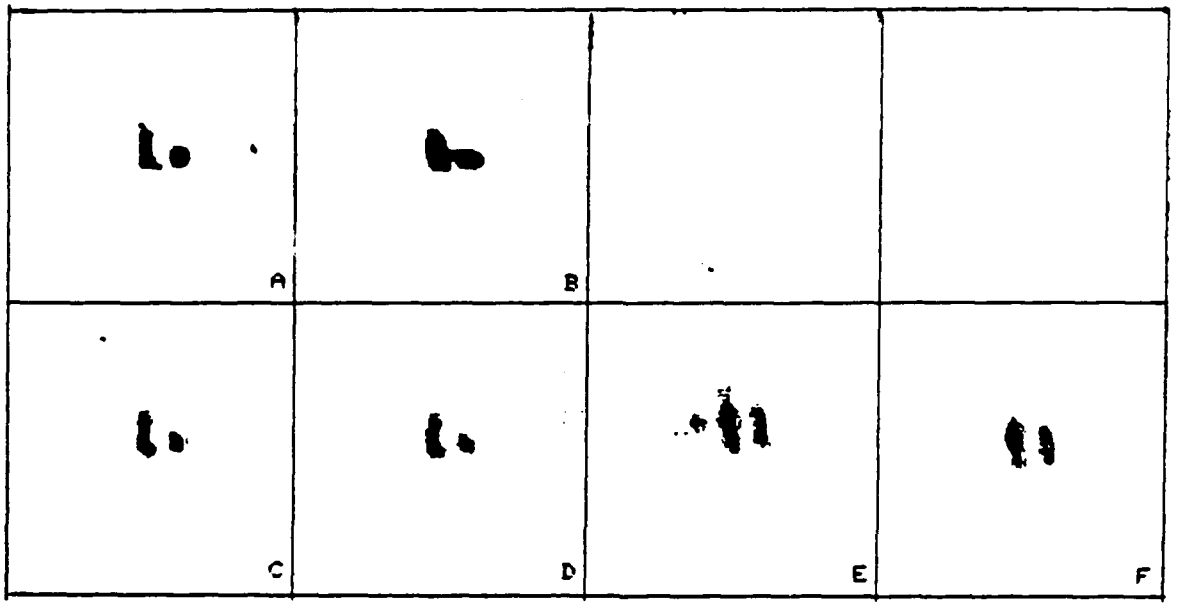


Fig. 11 Results with the positive-real, real constraint, and aberrator. (A) Reconstruction with truth data, (B) Mask for guess used to start phase retrieval (C) Reconstruction, no epoxy abberator, before positive-real constraint, (D) Reconstruction, no epoxy abberator, after positive-real constraint, (E) Reconstruction, epoxy abberator, before positive-real constraint, (F) Reconstruction, epoxy abberator, after positive-real constraint.

violated the nonnegativity constraints we discussed previously. Because of this we examined the effects of recovering the image with and without this constraint. The key results are shown in Fig. 11. The reconstructed truth data is shown in Fig. 11(A): the starting mask is shown in Fig. 11(B). Figures 11 (D) and (E) show the insignificant effect of the constraint on an object with no aberrator. Figures (E) and (F) show the effect on the image recovered with the epoxy aberrator data. The images recovered are extremely poor and it is not possible to discern the two parts of the object image or see any significance in what is recovered. We can conclude that the effect of the aberrator was quite significant and that the speckle intensity data was drastically altered. We can also conclude that enforcing the nonnegativity constraint was not effective in overcoming the effects of the aberrator. This suggests that we will have to examine the entire concept of the "thin" aberrator and come up with a model for what this means. It is obvious either the thickness of the aberrator was significant or there are significant amplitude transmission variations in the aberrator that were not visible in the cursory observations we made under room light.

VIII. CONCLUSIONS AND RECOMMENDATIONS

From the experimental results shown here we can reach the following conclusions:

(1) It is possible to conduct certain types of IC experiments with a fairly simple and economical setup using photographic film to carry out speckle recording and a lens to recover the autocorrelation of the object.

(2) We could not find evidence that glints improved resolution using conventional IC processing techniques; however, we did find results that suggest that glints might create dynamic range problems.

(3) We did not find the expected aberration correcting capability of IC using a thin epoxy aberrator.

We feel that the IC approach we have developed will make it easy for a larger number of researchers with a modest budget to participate in IC and we encourage its use for investigating certain issues.

We believe there is still a good possibility that glints can be helpful in improving image resolution and signal-to-noise. But this is not true using the present IC approach . We suggest instead, trying approaches that use the Fourier transform hologram information in the starting guess. Perhaps high pass filtering of the Fourier modulus can be used to emphasize this image information.

We were surprised by the aberrator results and feel it is necessary to put an effort into studying the abberator and its effect on the speckle directly. Although, there was neither time nor money available within the present contract to pursue this important area the aberration correction potential of IC alone is an important issue in many practical Air Force problems.

REFERENCES

1. Idell P.S.,Fienup J.R , and Goodman, "Image Synthesis from Nonimaged laser speckle patterns", Optics Letters, p858-860, Nov. 1987

2. Dainty J.C. (ed.), "Laser Speckle and Related Phenomena", 2nd edition, Topics in Applied Physics, Vol. 9, Springer, Berlin, 1982

3. Goodman J.W., "Introduction to Fourier Optics", McGraw Hill, p151, 1968

4. Gold Fisher L.I., "Autocorrelation Function and Power Spectral Density of Laser Produced Speckle Patterns", J. Opt. Soc. Am., Vol 55, p247-253, 1965

5. Bracewell, R. N., "The Fourier Transform and its Applications", McGraw Hill (San Francisco), p115, 1978

6. Fienup J.R., "Reconstruction of an Object From the modulus of it's Fourier Transform", Opt. Let., Vol 3, p 22-25, 1978

7. Fienup J.R., and Wackerman W.C., "Phase Retrieval Stagnation Problems and Solutions", J. Opt. Soc. Am. A, Vol 3, p 1897-1907, 1986

8. Knopp J. and Spielbusch B. "Experimental verification of Imaging Correlography" unpublished Aug. 1987

9. Personal conversation with Dave Homes, AFWL, Summer 1988

10. "Kodak TMAX-100 tech sheet"

11. Knopp J. and Becker M.F. "Virtual Fourier transform as an analytical tool in Fourier optics", Applied Opt., p1669, Vol 17, June, 1978

12. Personal communications with D.H. Williams, The Department of Electrical and Computer Engineering, The University of Texas at El-Paso.

FINAL REPORT NUMBER 49
REPORT NOT AVAILABLE AT THIS TIME
Dr. Barry McConnell
760-7MG-047

1987 USAF-UES SUMMER FACULTY RESEARCH PROGRAM/
GRADUATE STUDENT SUMMER SUPPORT PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the Universal Energy Systems, Inc.

RESEARCH INITIATION PROGRAM (R.I.P.) FINAL REPORT

INTERACTION OF LASERS WITH SUPERCONDUCTORS

Principal Investigator: Randall D. Peters, Ph.D.
Associate Professor of Physics
P. O. Box 4180
Texas Tech University
Lubbock, Texas 79409
(806) 742-3767

Research Sponsor: USAFWL/TAL
Kirtland AFB NM 87117-6008

USAF Technical Advisor: Dr. Patrick J. Vail

Date: 26 December 1988

USAF-UES Contract No: F49620-85-C-0013

R. I. P. FINAL REPORT

INTERACTION OF LASERS WITH SUPERCONDUCTORS

ABSTRACT

The high Tc superconductor $YBa_2Cu_3O_7$, with a transition temperature at about 95K, has been studied for target hardening purposes. In studies conducted at Texas Tech University during the course of this minigrant (January-December, 1988), it was found that the reflectance of the material is not appreciably temperature dependent in the range from 77K to 300K, for wavelengths between 1μm and 20 μm. Thus it is concluded that the bandgap must lie beyond 20μm, probably in the far infrared or microwave region of the electromagnetic spectrum. Therefore the superconducter is not expected to be useful for hardening against laser radiation.

I. INTRODUCTION

For this minigrant, it was proposed to study the optical, thermal, and electrical properties of high Tc superconducting samples at several intensities and wavelengths as the sample temperature was varied through the transition temperature from normal to superconducting, Tc. Reasons for the study are indicated in the attached proposal; which also, for background purposes, refers to the author's summer faculty fellowship final report [1]. Thus a program was initiated at Texas Tech University whereby: (i) temperature of the samples could be controlled, and (ii) reflectance could be measured under temperature controlled conditions. The measurements were planned around both (a) a strong fixed wavelength source of 1.06μm using a Nd:YAG laser and (b) a weak variable wavelength source. The variable source was a blackbody in which wavelength was selected using a scanning monochromator. Measurements were not performed with the YAG laser, primarily because of a difficulty beyond our control--our laboratory had to be moved from the science building to chemistry for purpose of asbestos abatement.

It was expected that the integration of components (cryogenic, vacuum, and optics) would be a difficult challenge. Most of the subsystems were to be "homemade" since the physics department lacked commercial subsystems to do the job. This was not considered a futile exercise because the minigrant was understood to be, among other things, a vehicle for student learning. We initially believed that temperature control would be necessary, although later results showed that less time should have been invested in this area. Likewise, a large amount of time was invested in building a blackbody/monochromator source. All of these activities would have taken a dramatically different course if a commercial spectrometer had been available for use in January rather than in December. A Perkin-Elmer Fourier Transform Infrared (FTIR) spectrometer was acquired late in the year by a colleague through a Texas Advanced Technology grant award. Through Dr. Gangopadhyay's permitting us free access to this valuable instrument (model 1600 series), we were able to recover from what would otherwise have been an unsatisfactory performance. We can testify from first hand experience that such an instrument is far superior to the blackbody/monochromator. One should strictly avoid the latter unless he has only free time and not the capital required to purchase an FTIR at the minimum price tag of about $15K.

II. TEMPERATURE CONTOLLER

Although a good controller and cryostat were later available on loan

from Dr. Lichti (some of his Welch grant equipment, Lakeshore Cryotronics Inc., DR80C controller and Janis model DT cryostat), we initially were forced into building our own. This we did by using duty cycle control via an inexpensive microcomputer, according to Ref. [3]. Miss Mythili Sankaran constructed this controller using the Radio Shack Color Computer II with a platinum resistance thermometer. Toward the end of her work the commercial unit became available, and so the homemade unit was not used for the data here reported. Information concerning her system is included in Appendix I. A paper on the subject was given at the Nov. meeting of the Texas Section of APS/AAPT [4].

III. BLACKBODY/MONOCHROMATOR

Data in the region from 1μm to 3μm were collected using a blackbody consisting of an electrically heated carbon rod in a water cooled brass cavity. This unit is essentially the Rao source described in Ref. [5], and it was left over in the department from the work of Ref. [6]. A massive step down transformer provides about 200 amperes at 20 volts from a 110V line source. Its output versus wavelength is shown in Figure 1 in the range from 700 nm to just beyond 1.5 μm. The secondary hump between 1120 nm and 1500 nm is not from the source; it corresponds to 2nd order (m=2 rather than ideal m=1 only) 650 nm radiation. Order blocking filters were not available for the blazed Oriel grating. The grating was purchased for this work, as part of a stepping motor controlled monochromator (Oriel model no. 77250 anb gratings 77298-77303). The purchase was possible, along with an integrating sphere and pyroelectric detector (as well as the planned but unused Nd:YAG laser) using startup monies provided to the principal investigator by the university. Not enough money was available to purchase the filters. Control of the stepper motor is via an APPLE IIE computer, so that wavelength range and speed of scan are readily managed. The pyroelectric detector used in this case (model 7080) was good for checking the spectral characteristics of the source, but its NEFD is too poor for reflectance measurements on the samples using an integrating sphere. The total reflectance is evidently only of the order of a few percent and thus the pyroelectric detector response was very noisy. Therefore a PbS element operating as a photoconductor was used for the reflectance measurements. The 2nd order 650 nm radiation is no problem in this case because PbS responds only in the 1μm to 3μm region. The detector is also from Almond's work [6], and it was necessary to cool it to 77K for proper operation. It was operated in a bridge circuit driven with about 40 vdc,

the three other resistors of the bridge having comparable resistance to the PbS element at 77K. Figure 2 shows the response of the photoconductor to the blackbody source. The source of the dip around 1.9 μm is not known. This data was taken by focusing radiation from the source onto the entrance aperture of the integrating sphere (gold coated infrared type of 3 in. dia. from Oriel). The exit aperture could be mounted only in close proximity to the window of the PbS detector. This window was necessary because of the homemade dewar which houses the element. A reflectance measurement from one of the high Tc samples at room temperature is shown in Figure 3.

A vacuum housing was built to hold the sample. This was considered necessary to avoid the formation of frost in going below the transition temperature. As had been expected, the integration of all these components proved unwieldy. In spite of numerous attempts to calibrate the system, it was determined that the monochromator measurements could not be trusted for absolute reflectance values. Thus Fig. 3 is in relative units. This data does overlap the FTIR measurements, however. It can thus be calibrated with respect to the specular part, which is the only component of reflectance the FTIR instrument was capable of measuring.

IV. REFLECTANCE RESULTS

There were two different sample types used for this work. One was a film material on a thin sapphire substrate. The first sample of this type which we studied went bad after a short while. Probably, the Oxygen stoichiometry became unacceptable with time. The second sample of the same type proved to remain good for the duration of the studies. The other sample type (only one studied) was a bulk superconductor. These shall later be referred to as the thin and thick sample, respectively. In all cases the samples were of the Yttrium/Barium type. They were manufactured at Sandia and shipped to us via the weapons laboratory.

The monochromator result is shown in Fig. 3, already mentioned. It was taken using the thin sample. We had planned to obtain some low temperature values for both samples with this setup, but decided it was not worth the effort. The FTIR results are much easier to obtain, and they almost cover the 1.2 to 2.8 range of the monochromator. The fact that there is little difference between the reflectance above and below Tc in the FTIR cases insures that nothing significant is likely to be seen by repeating the Fig. 3 run with the sample at 77K.

Fig. 4 shows the thin sample room temperature reflectance (specular component) in the range from 2.3 μm to 20 μm, as determined using the

FTIR spectrometer. Evidently the absorptance for these superconducting materials is fairly high. That was noted by looking at the reflectance from the sapphire substrate (simply turning the sample over). It is to be inferred from the fact that substrate reflectance reaches 70.5% at 15μm, whereas Fig. 4 shows no appreciable rise at this wavelength.

The thick sample reflectance at room temperature is shown in Fig. 6. It is seen that the optical properties of the two are rather similar. It was also found, as noted in later figures, that the resistivity vs temperature plots were very similar.

LOW TEMPERATURE REFLECTANCE

Figs. 7 and 8 show the thin and thick sample reflectances, respectively, in the neighborhood of 77K. These curves were generated by immersing the samples in liquid nitrogen, and then simply placing them on the horizontal surface of the reflectance fixture of the spectrometer. The scan time, at a resolution of 8 cm^{-1} was only a couple of seconds, assuring that the sample did not rise above Tc during the course of the measurement. Also, the amount of ice formation on the samples during this time was insignificant. Comparing Figs. 7/8 with Figs. 4/6 illustrate the negative results of this study. There is insignificant change in the reflectance above and below the transition temperature, as also noted in a Phys. Rev. paper [2]. There are two possible conclusions. Either (i) the superconductivity is not a surface as well as bulk phenomenon, and penetration to the region of superconduction is highly attenating, or (ii) the energy with which the electrons are bound (Cooper pairing) corresponds to low energy photons in the far infrared. It may be that both of these factors are at work. It has been said, however, that Los Alamos microwave experiments have demonstrated Q improvements when cavities were lined with the material. Thus it appears that the bandgap argument is the more likely one.

V. RESISTANCE MEASUREMENTS

The resistance vs temperature was obtained for both the thin and thick samples using the commercial cryostat. The results are shown in Figs. 9 and 10 for the thin and thick samples respectively. It is seen that the samples are indeed superconducting with a transition temperature in the neighborhood of 95K. This was verified both before and after the reflectance measurements.

VI. CONCLUSIONS

Looking at the minigrant program in retrospect, it is obvious that many of the efforts at Texas Tech University were misdirected. The principal investigator takes sole responsibility for that. This is probably not surprising, however, considering the nature of the program and the constraints of operation. The $20K award is too small to permit capital equipment acquisition, and thus resulted in a "shoestring" operation. What was learned is nevertheless worthwhile. The reflectance data we obtained, coupled with that from Ref. [2] show that the bandgap of $YBa_2Cu_3O_7$ must correspond to a wavelength in excess of 100μm. This means that any hardening capability for the material is possible only for microwave, rather than laser radiation.

The program did provide a genuine learning experience which is beneficial to Texas Tech University. Moreover, it is hoped that our experience in spectral data collection could assist others as they plan future work; i.e., that an FTIR is well worth its purchase price. Although the acquisition cost is initially greater than that of older instruments, the time that is later saved by its use makes the "upfront" cost justifiable.

REFERENCES

1. Randall D. Peters, "Momentum Transfer and Mass Loss for a C.W. Laser Irradiated Target", 1987 USAF-UES Summer Faculty Research Program/Graduate Student Summer Support Program, 6 Aug. 1987.

2. D. A. Bonn, J. E. Greedan, C. V. Stager, and T. Timusk, "Far-Infrared Conductivity of the High-Tc Superconductor $YBa_2Cu_3O_7$", Phys. Rev. Lett. 58, no. 21, pg. 2249, 25 May 1987.

3. R. D. Peters, "Experimental Computational Physics using an Inexpensive Microcomputer", Computers in Phys. 2, no 4, 68, July/Aug 1988.

4. M. Sankaran and R. Peters, "Inexpensive Microcomputerized Proportional Temperature Controller", Amer. Phys. Soc./Amer. Assoc. Phys. Teach. Tex. Sec. Mtg., Lubbock, TX, Nov. 1988.

5. K. N. Rao, C. J. Humphrey and D. H. Rank, Wavelength Standards in the Infrared (Academic Press, New York, 1966), pg. 146.

6. William H. Almond, "Infrared Analysis of the Non-Degenerate Symmetric Stretch Vibration of Methyl Alcohol", PhD dissertation , Texas Tech University, Aug. 1972.

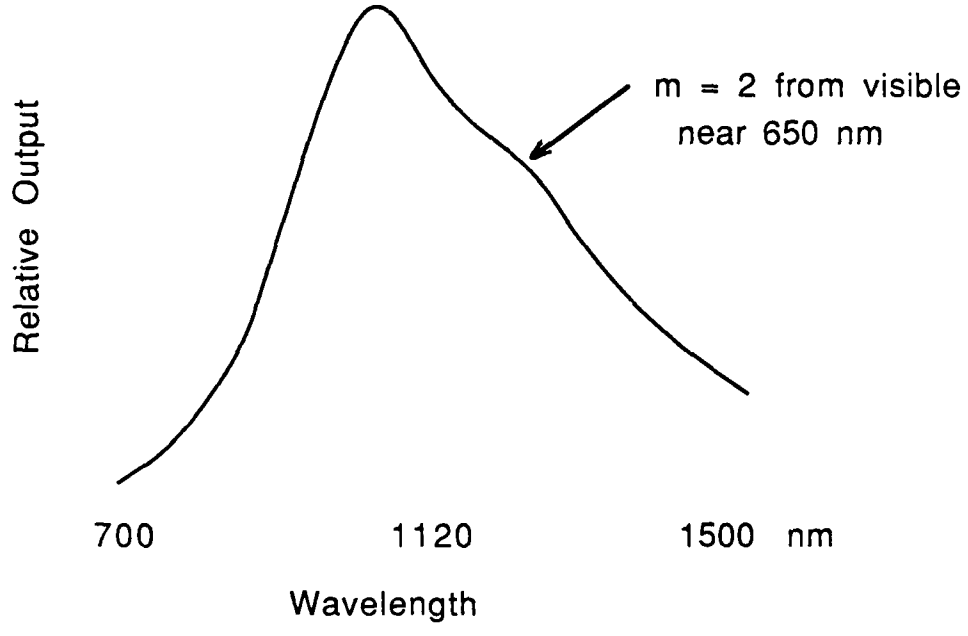


Figure 1 Spectral Characteristics of the Blackbody Source

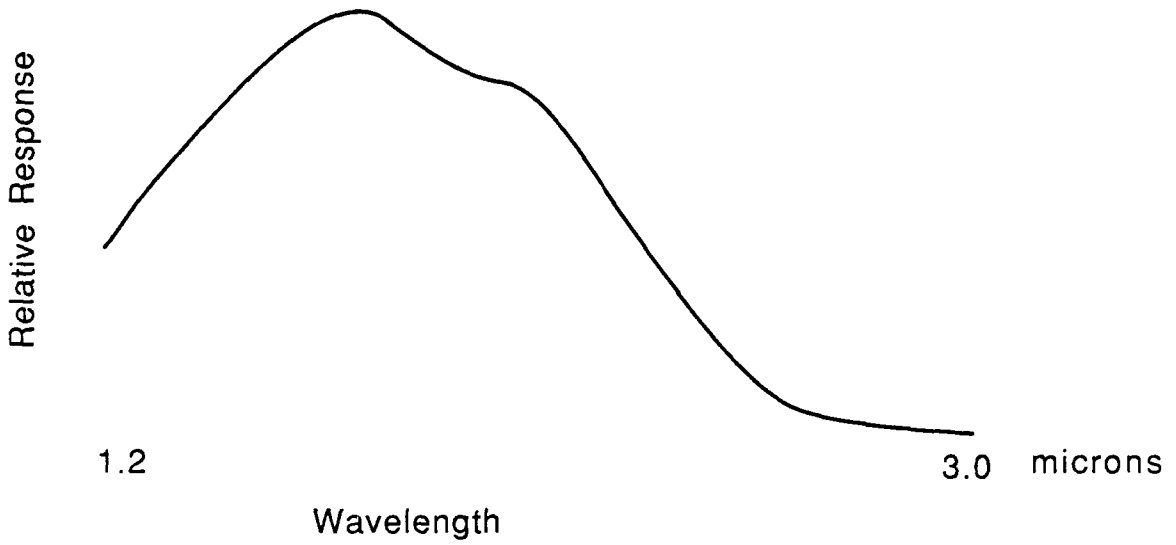


Figure 2 Response of PbS Detector to Blackbody Source

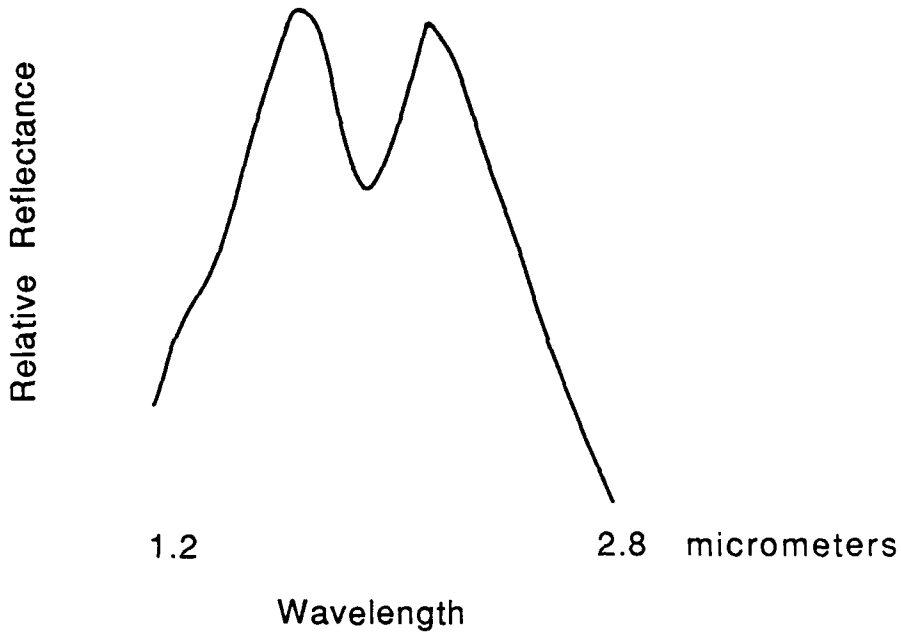


Figure 3 Room Temperature Reflectance from Thin Sample obtained with Monochromator and Integrating Sphere

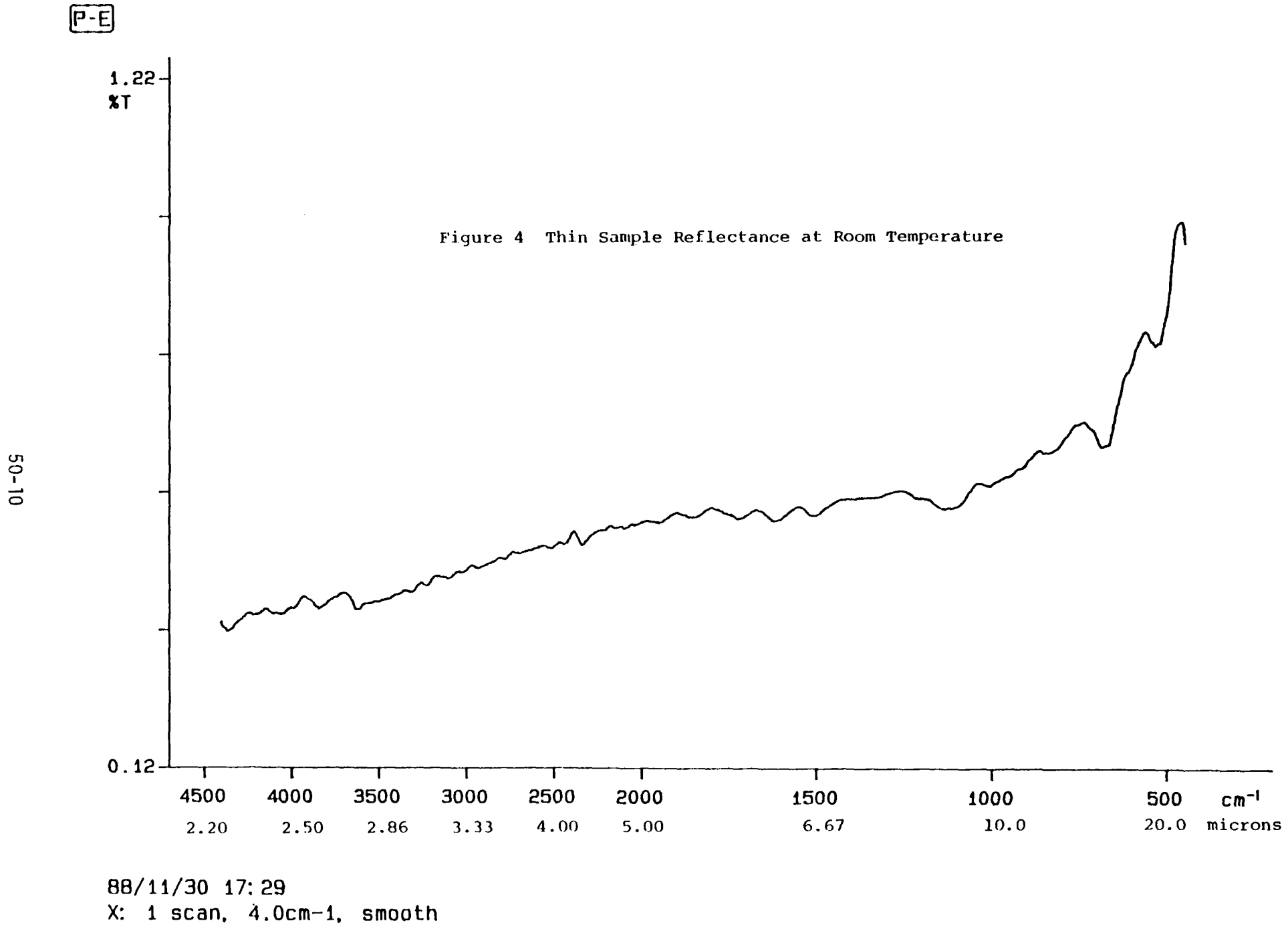


Figure 4 Thin Sample Reflectance at Room Temperature

P-E

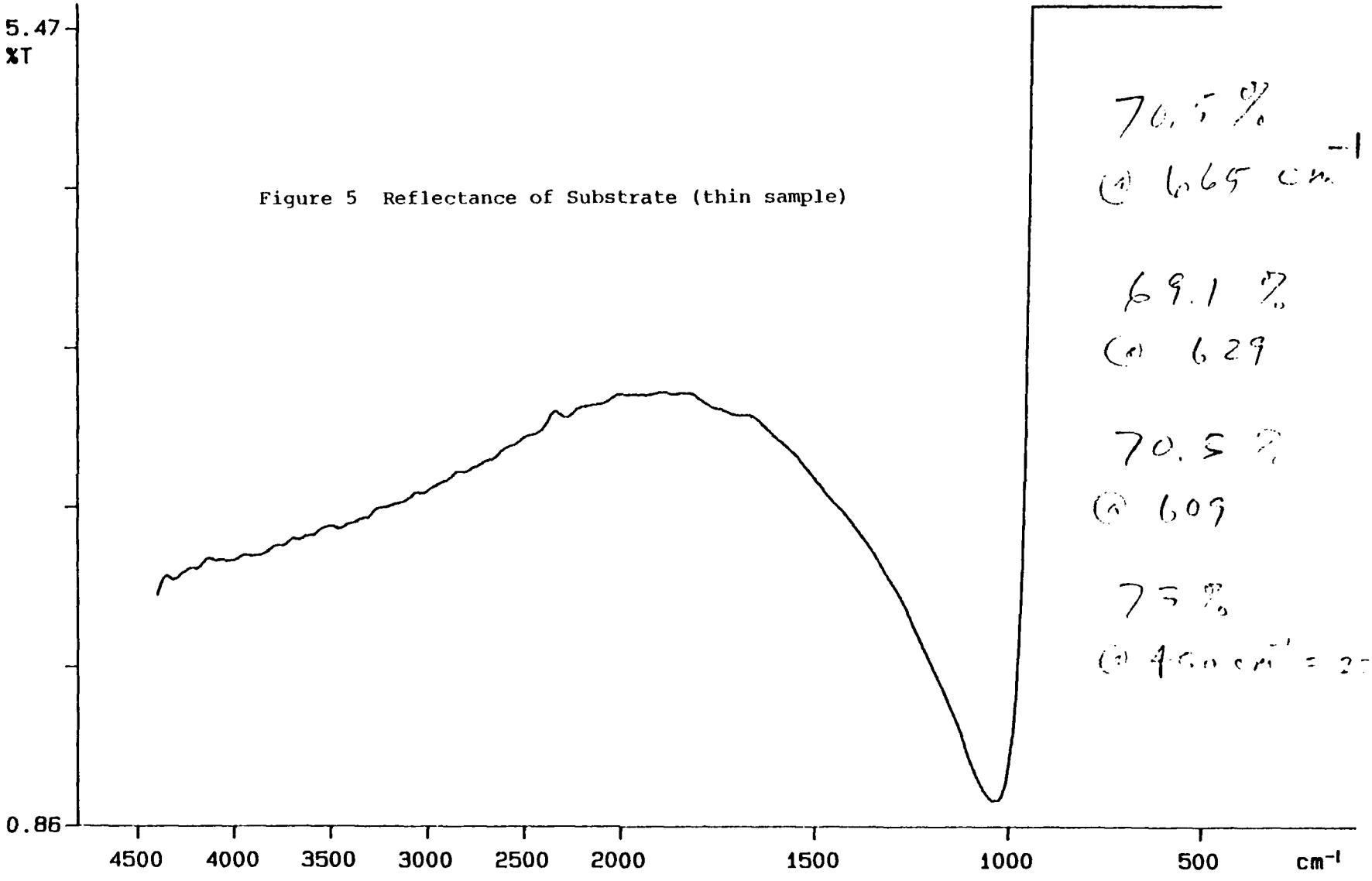


Figure 5 Reflectance of Substrate (thin sample)

70.5 %
@ 665 cm-1

69.1 %
@ 629

70.5 %
@ 609

75 %
@ 4[illegible] cm-1 = 2[illegible]

50-11

88/11/30 17:48
Y: 1 scan, 4.0cm-1, smooth

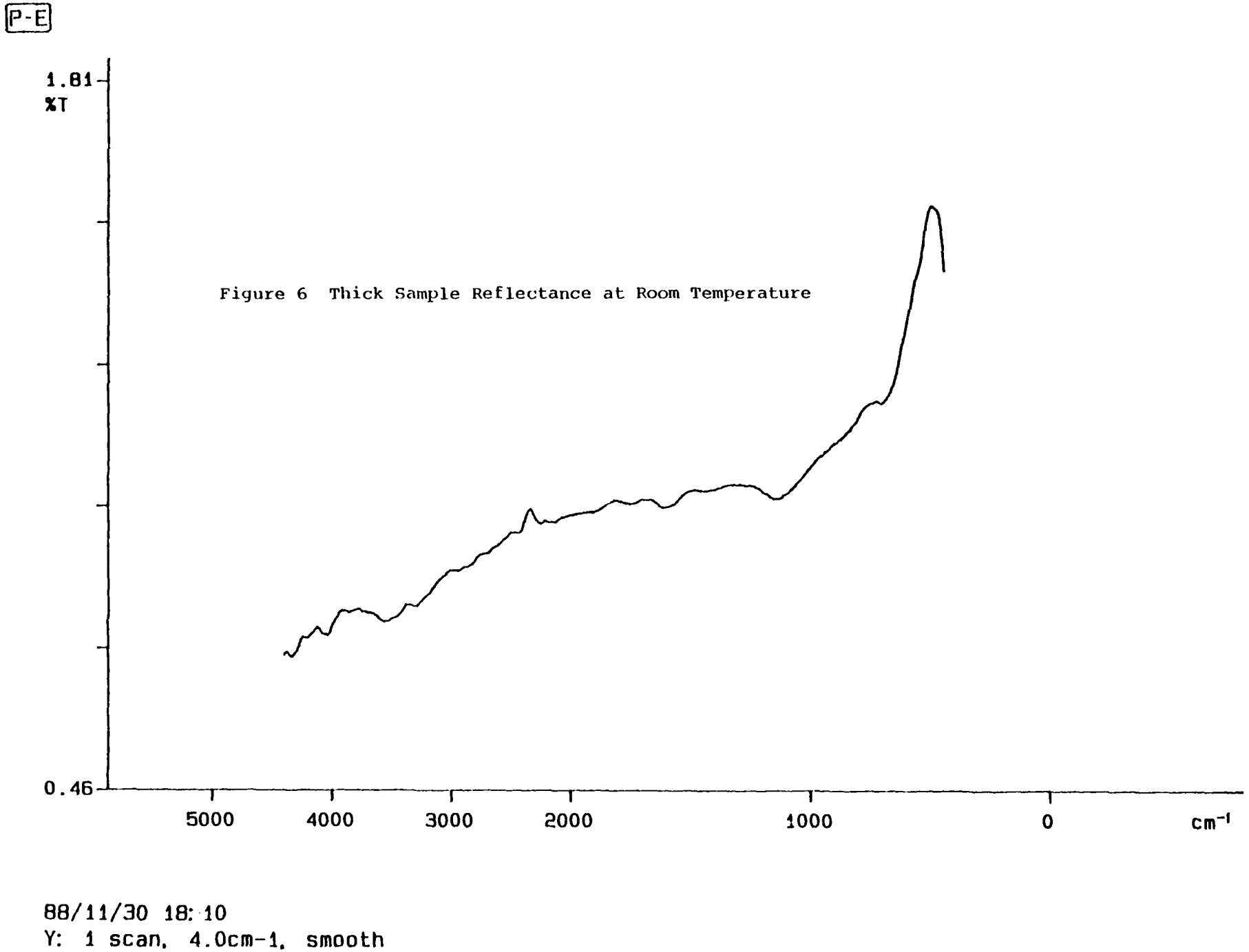


Figure 6 Thick Sample Reflectance at Room Temperature

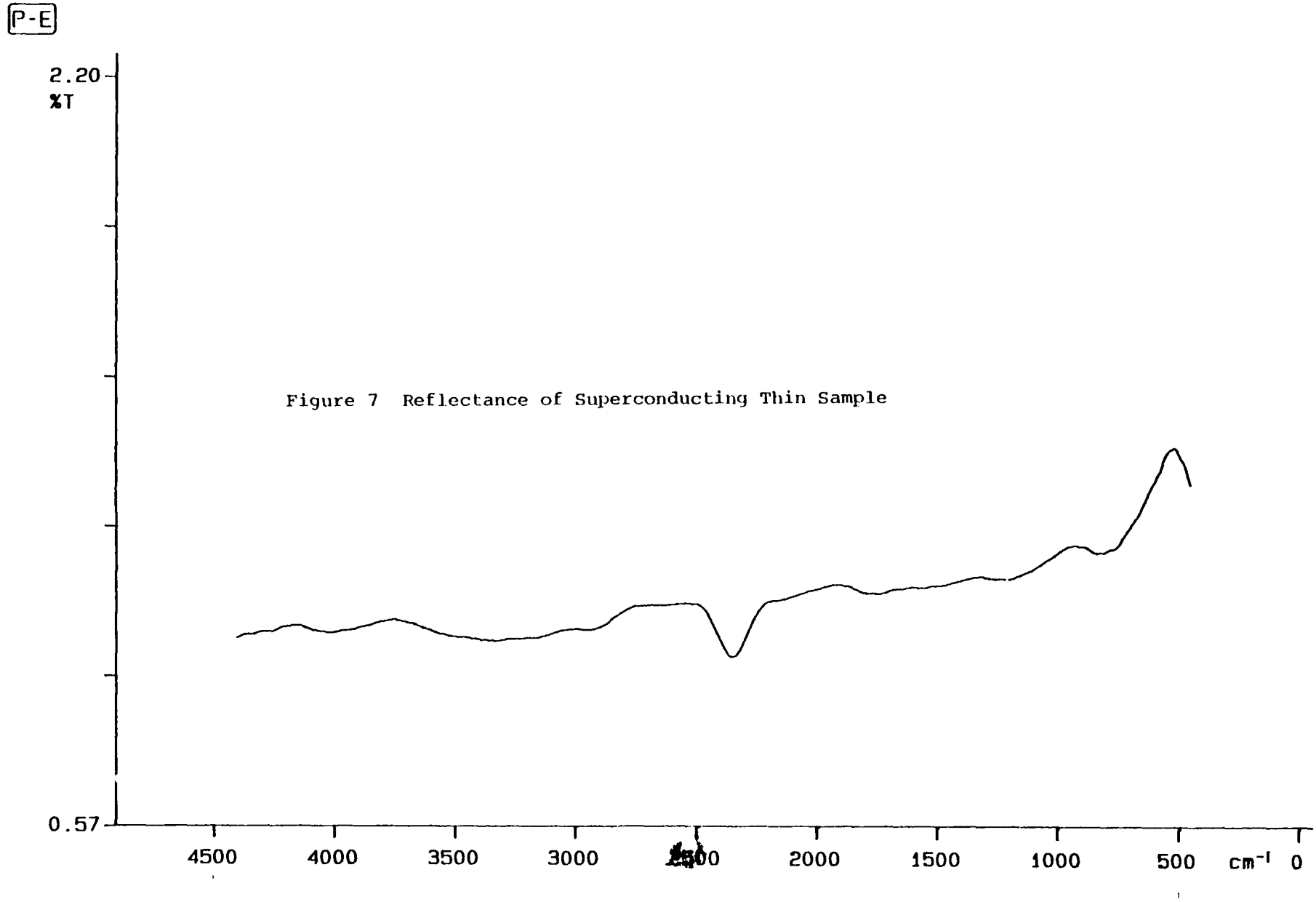


Figure 7 Reflectance of Superconducting Thin Sample

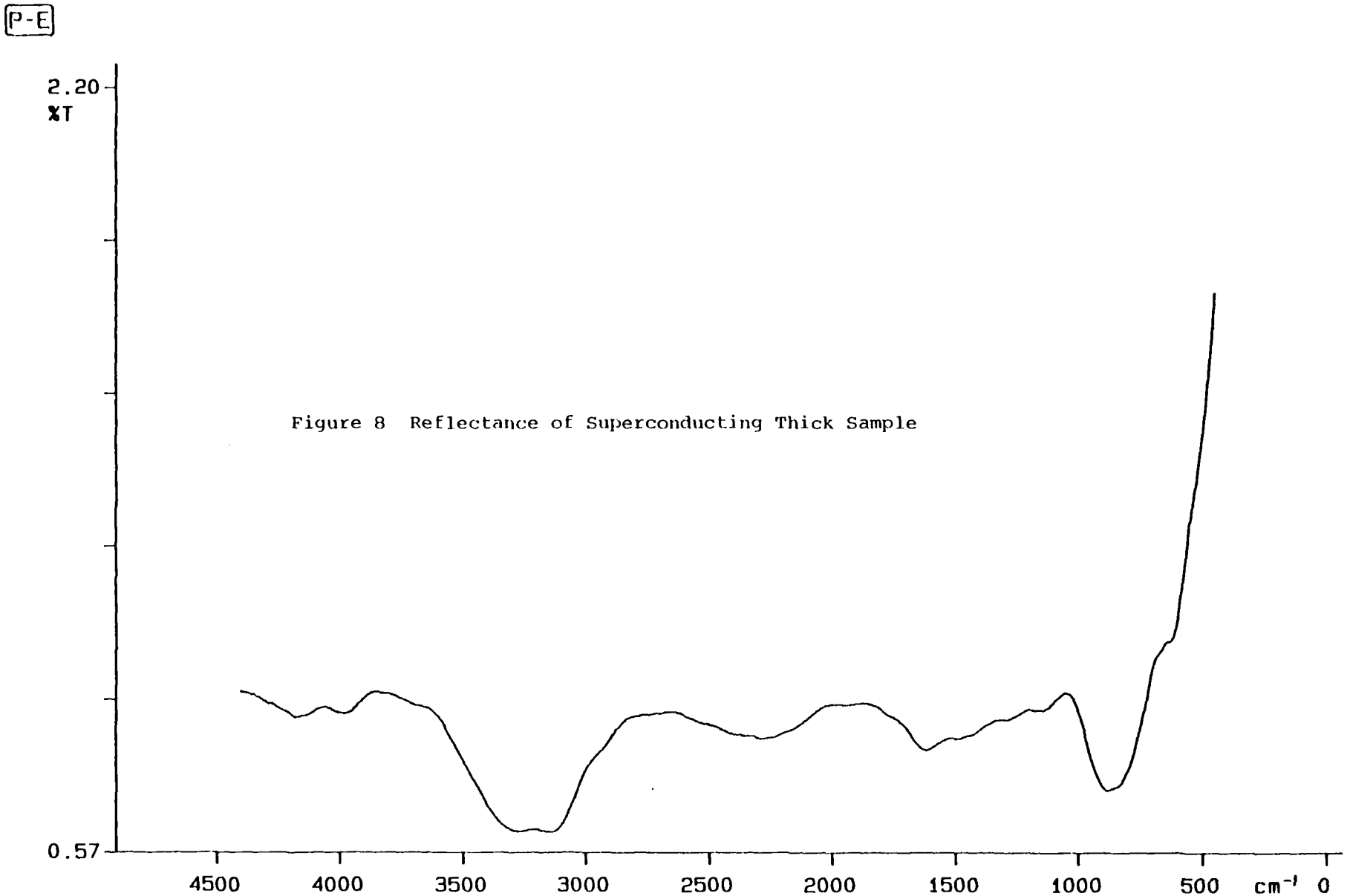


Figure 8 Reflectance of Superconducting Thick Sample

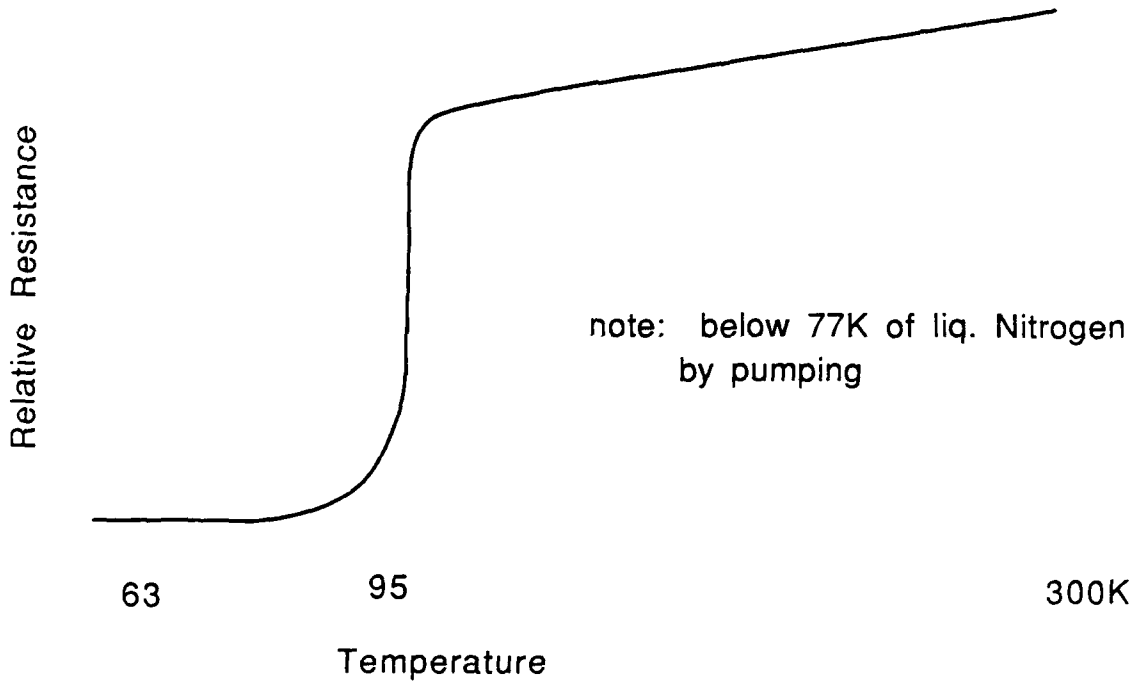


Figure 9 Resistance vs Temperature for Thin Sample

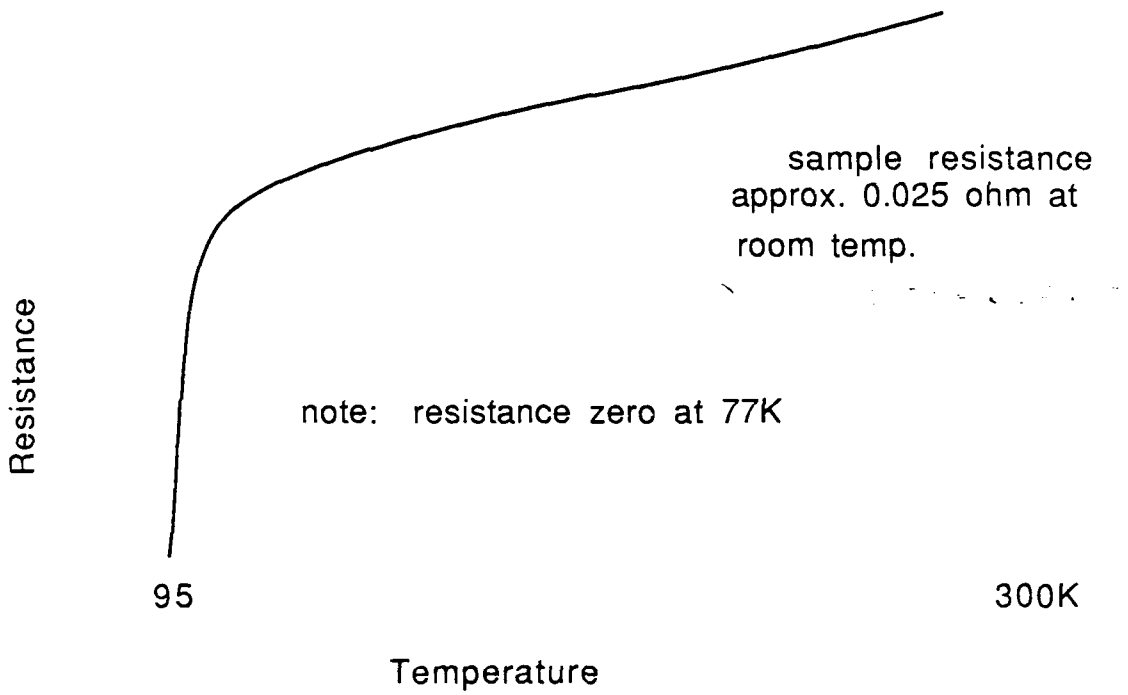


Figure 10 Resistance of Thick Sample vs Temperature

Appendices can be obtained from
Universal Energy Systems, Inc.

FINAL REPORT

"Three Dimensional Thermal Conduction Effects in High Power CW Laser Target Plates"

By: Martin Andrew Shadday Jr.
Asst. Prof. of Mech. Engr.
University of South Carolina

ABSTRACT

The Air Force Weapons Laboratory has developed a simple and inexpensive technique for measuring the spatial intensity distributions in the cross-sections of high power continuous wave lasers. The intensity pattern on the front surface of a thin metal target plate is determined from the response of temperature sensitive paint on the back surface, assuming one-dimensional heat transfer through the plate. Thermal conduction through the target plate has been modeled numerically, and various spatial intensity distributions have been run to quantify the resolution of the target plate measurement technique. For narrow features in the laser beam intensity distribution. the assumption of one-dimensional heat conduction can result in significant errors.

NOMENCLATURE

A	Area
C_p	Specific heat
h_{rad}	Radiation heat transfer coefficient
h_{conv}	Convection heat transfer coefficient
k	Thermal conductivity
L	Target plate thickness
Nu	Nusselt number
q	Intensity of thermal radiation (w/cm^2)
q_o	Intensity of the uniform background
q_p	Peak intensity
T	Temperature
T_f	Fluid temperature
T_s	Surface temperature
t	Time
t^*	Non-dimensional time
α	Thermal diffusivity
α_o	Reference thermal diffusivity
ϵ	Emissivity
ρ	Density
σ	Stefan-Boltzmann constant

INTRODUCTION

The spatial intensity distribution in the beam cross-section is a fundamental characteristic of a laser. The measurement of intensity distributions in high power lasers is difficult because of the destructive nature of the high power laser radiation. Data must necessarily be collected in a short period of time, and consequently thermal measurement methods are most appropriate. The Air Force Weapons Laboratory has developed a simple and inexpensive thermal method for measuring intensity distributions in the beam cross-sections of high power continuous wave lasers, Lamar (1). The beam intensity distribution is determined from the transient temperature response of the back side of a thin metal plate, illuminated on the front by the laser beam. Thermal conduction through the metal plate slows the rate at which data must be collected. The front surface of the metal target plate heats up very quickly. Too quickly for the two-dimensional transient temperature distributions to be accurately measured, at the intensities encountered with high power continuous wave lasers. The rear surface of the target plate heats up much slower than the front, considerably simplifying data collection.

The spatial intensity distribution in the laser beam hitting the front surface of a target plate is determined

from the transient temperature on the back surface of the target plate, assuming one-dimensional thermal conduction through the plate. This assumption is necessary because of the manner in which transient temperature data for the back surface of the target plate is collected, and it considerably simplifies the solution of the inverse heat transfer problem. The back surface of the metal target plate is coated with temperature sensitive paint, and the growth of isotherms on the painted surface is recorded by a high speed movie camera. At a specified elapsed time, an isotherm on the rear surface of the target plate uniquely corresponds to an isointensity line of absorbed laser radiation on the front surface, only if the thermal conduction through the plate is one-dimensional. The transient temperature data requirements for the rear surface of the target plate are greater if multi-dimensional thermal conduction is to be accounted for.

Unless the spatial intensity distribution in the laser beam is uniform, the thermal conduction through the target plate is not truly one-dimensional. The purpose of this investigation is to quantify the measurement errors associated with the assumption of one-dimensional thermal conduction in the target plate laser intensity measurement technique. This is done by comparing the results of a three-dimensional heat conduction numerical model of the target plate with a numerical model of the target plate

with one-dimensional heat conduction. The capability of the target plate laser intensity measurement technique to resolve narrow intensity spikes in the beam cross-section is determined. Sharp intensity gradients in the laser beam cross-section lead to sharp thermal gradients in the target plate in directions parallel to the plate surfaces. As a consequence, thermal conduction in these directions can be important, and the measurement technique would predict intensity spikes with a lower peak intensity than actually exists. The ability of the target plate measurement technique to resolve adjacent narrow intensity spikes is also impaired by the assumption of one-dimensional thermal conduction through the target plate.

LASER INTENSITY MEASUREMENT TECHNIQUE

The target plate laser intensity measurement technique, developed by the Air Force Weapons Laboratory, measures spatial intensity distributions in the beam cross-sections of high power continuous wave lasers. Intensity measurements are made by directing the laser beam on to a thin metal target plate. The rear surface of the target plate is coated with temperature sensitive paint, and the transient thermal response of the paint is recorded by a high speed movie camera. The camera records the growth in time of an isotherm on the rear surface of the target plate, at the phase change temperature of the temperature sensitive paint. The thermal conduction through the target plate is assumed to be one-dimensional, and therefore there is a direct correlation between an isotherm on the back surface and an isointensity line of absorbed energy on the front surface. The assumption of one-dimensional heat conduction is necessary in order to solve the inverse heat transfer problem, with the transient temperature data collected.

The target plate is made of stainless steel, nickle, or copper depending upon the average intensity of the laser. With laser intensities near the upper end of the intensity range, copper is used because of its high thermal conductivity. It is important that the temperature sensitive paint on the rear surface of the target plate respond to the laser heating

before melting occurs on the front surface of the target plate. The front surface of the target plate is highly polished and in some cases gold coated in order to reduce its absorptivity.

The spatial intensity distribution in the laser beam is assumed to be constant over the exposure time of the target plate. For a laser with a time varying intensity pattern, the target plate measurement technique gives time integrated results. Depending on the intensity of the laser radiation, exposure times vary from fifteen milliseconds to one-half second. A fast shutter opens to expose the target plate to the laser beam. This avoids the beam transients at the onset of lasing, and provides a precise start time for the exposure of the target plate. The laser beam cross-section is typically circular, with a diameter on the order of six inches. Laser intensities are in the range of ten to one hundred kw/cm^2.

PROBLEM DESCRIPTION

The problem considered in this investigation is the transient thermal conduction in a thin metal plate. exposed on one side to laser radiation. The intensity of the incident laser radiation varies spatially, and it is assumed to be constant with respect to time. The transient thermal conduction in the metal plate is modeled numerically, and the output of the model is the transient temperature distribution on the back surface of the metal plate. This is the experimental data determined from the response of the temperature sensitive paint on the back surface of the target plate in the laser intensity measurement technique.

The intensity distribution of the incident laser radiation consists of several spikes superimposed on a background radiation field with uniform intensity. The lateral extent of the uniform intensity background radiation field, beyond the spatially varying intensity pattern, is sufficient to justify treating the outer edge of the metal plate as an insulated boundary. Figure (1) is a schematic of the problem.

The metal plate undergoes large temperature changes, and the temperature dependence of the thermophysical properties is accounted for in the numerical model. Thermal radiation and natural convection losses on the front and back surfaces of the target plate are also included in the model.

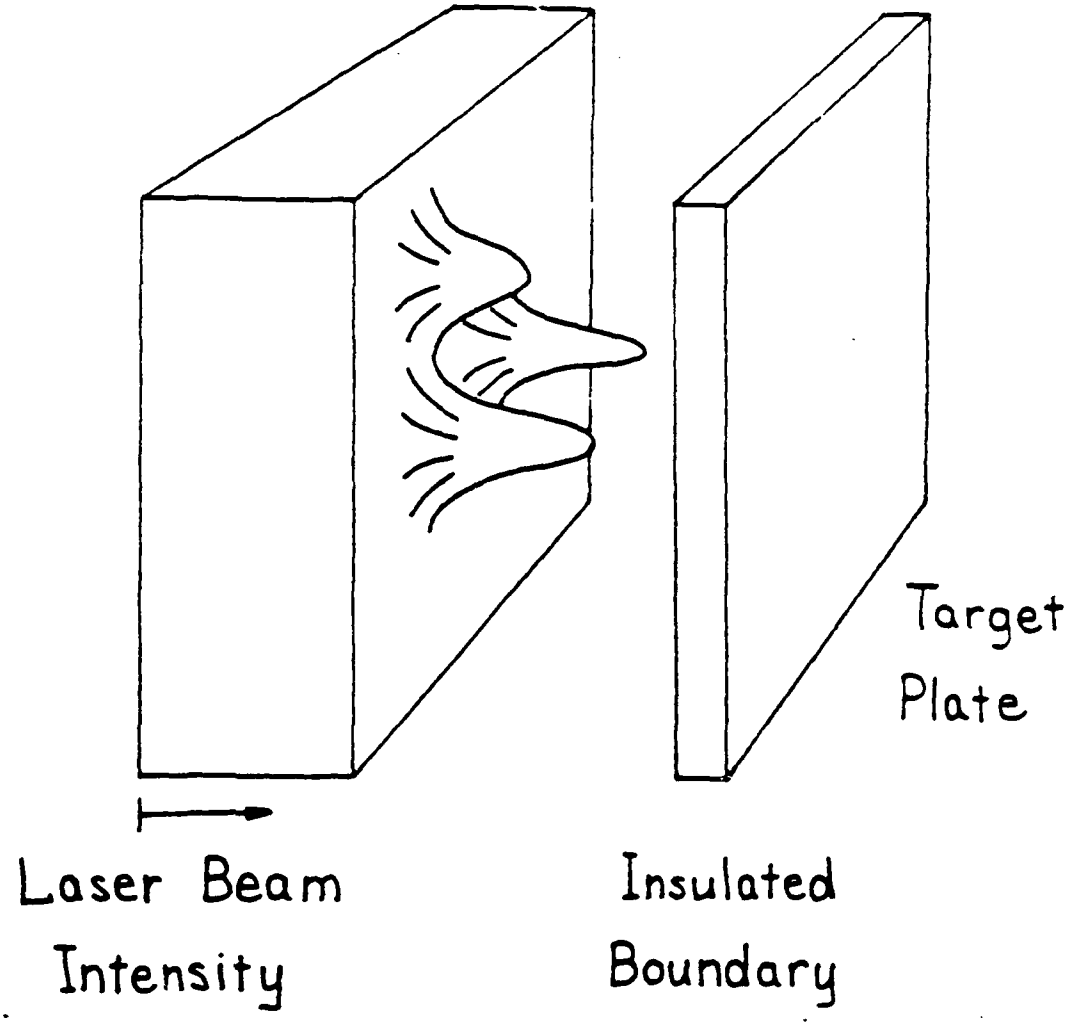


Figure 1

Schematic of the target plate and the incident laser radiation with an arbitrary spatial intensity distribution.

Two numerical models of transient conduction in the target plate were used, one a three-dimensional conduction model and the other that allowed conduction only in the direction normal to the plate surfaces. The magnitudes of the errors associated with the assumption of one-dimensional conduction in the target plate were determined by comparing the results of the two numerical models for a specified input laser intensity distribution.

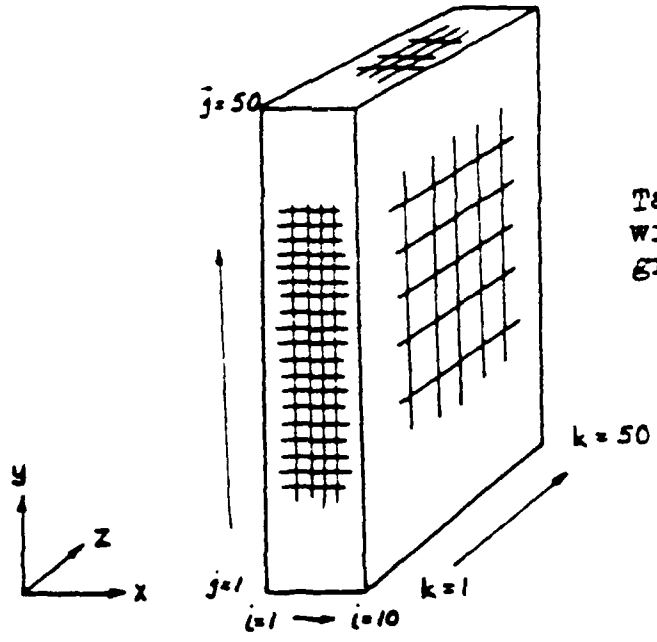
NUMERICAL MODEL

The numerical model of transient heat conduction through the target plate is a three-dimensional, partially implicit finite-difference model. The computational mesh is uniform in size, and it consists of ten nodes distributed through the thickness of the plate and fifty nodes in each of the orthogonal directions parallel to the plate surfaces.

Thermal conduction through the target plate is governed by the Fourier heat-conduction equations. Finite-difference analogs of these equations, for each of the nodes in the computational mesh, are derived from application of the conservation of energy principle to each of the incremental control volumes:

$$\rho C_p (Vol) \frac{\Delta T}{\Delta t} = \text{net heat rate in} \qquad (1)$$

Figure (2) is a schematic of an interior node and the associated incremental control volume. Application of equation (1) to the control volume results in an expression for the nodal temperature, at a new time level, in terms of the six neighboring nodal temperatures. In the two directions parallel to the plate surfaces, the finite-difference equations are explicit, that is the neighboring nodal temperatures are at the old time level. In the direction normal to the plate surfaces, the finite-difference



Target Plate
with the computational
grid

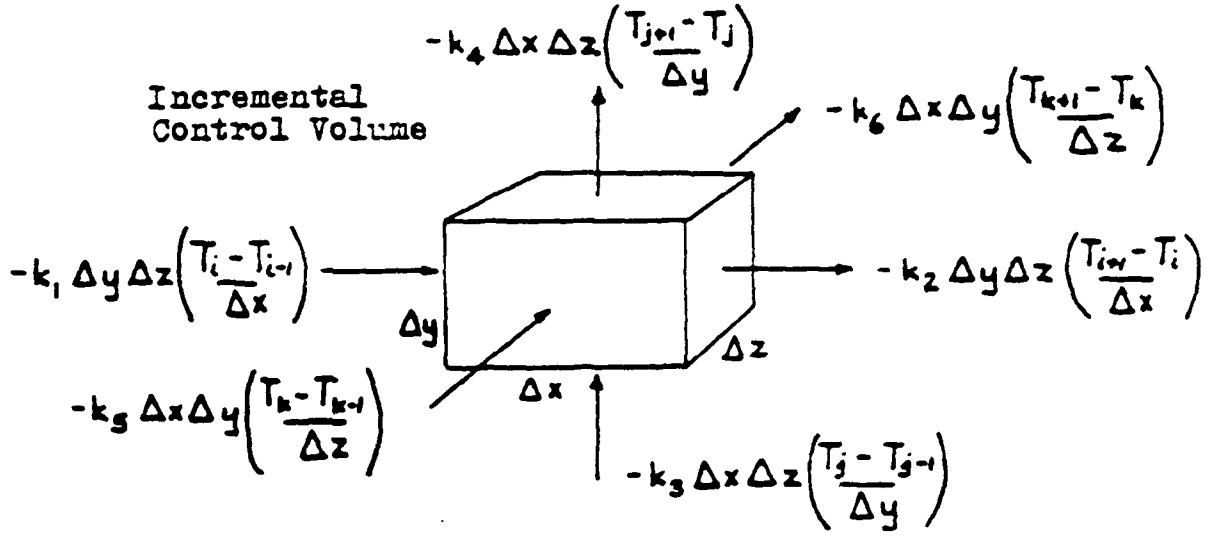


Figure 2

Schematic of the computational grid and an interior nodal control volume with the associated heat transfer rates crossing the control surface.

equations are implicit, that is the two neighboring nodal temperatures are at the new time level:

$$F\left(T_{ijk}^{n+1}, T_{i+1jk}^{n+1}, T_{i-1jk}^{n+1}\right) = G\left(T_{ijk}^{n}, T_{ij+1k}^{n}, T_{ij-1k}^{n}, T_{ijk+1}^{n}, T_{ijk-1}^{n}\right) \quad (2)$$

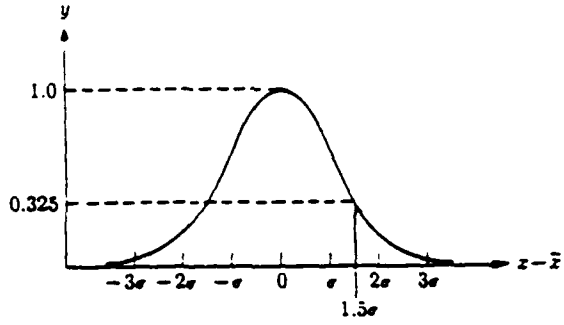
The finite-difference equations in the direction through the thickness of the target plate are coupled due to their implicit formulation, and they must be solved simultaneously. The ten simultaneous equations for a row of nodes, distributed through the thickness of the plate, have a tri-diagonal coefficient matrix:

$$\begin{bmatrix} & & & & 0 \\ & & & & \\ & & & & \\ & & & & \\ 0 & & & & \end{bmatrix} \begin{Bmatrix} T_1^{n+1} \\ T_2^{n+1} \\ \cdot \\ \cdot \\ \cdot \\ T_{10}^{n+1} \end{Bmatrix} = \begin{Bmatrix} \quad \\ \quad \\ \quad \\ \quad \\ \quad \\ \quad \end{Bmatrix} \quad (3)$$

This matrix is readily inverted, yielding the updated temperatures for the row along which the nodal equations are implicit. In a single timestep, the nodal temperatures in each of the rows normal to the front surface of the plate are updated sucessively by sweeping through the grid

in the two directions parallel to the front surface of the plate. Successive sweeps of the grid take place until the desired elapsed time has occurred.

The boundary conditions for the numerical model are convection and radiation boundary conditions on the front and back surfaces of the plate, and the four edges are adiabatic. The input for the model is laser radiation with an arbitrary spatial intensity distribution, as shown in figure (1). The intensity pattern of the laser beam consists of several spikes superimposed on a background radiation field with uniform intensity. The spikes are axi-symmetric and the shapes of their profiles are normal distributions:



$$y = e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \qquad (4)$$

The spike width is normalized with respect to the plate thickness, and it is defined where the amplitude is 32.5 percent of the peak amplitude. The absorptivity of the front surface of the target plate is assumed to be constant

at five percent.

Radiation and convection losses from the front surface of the target plate are included in the model. Average radiation and natural convection heat transfer coefficients are computed for the front surface of the plate each timestep, and local losses from each node are determined from Newton's law of cooling:

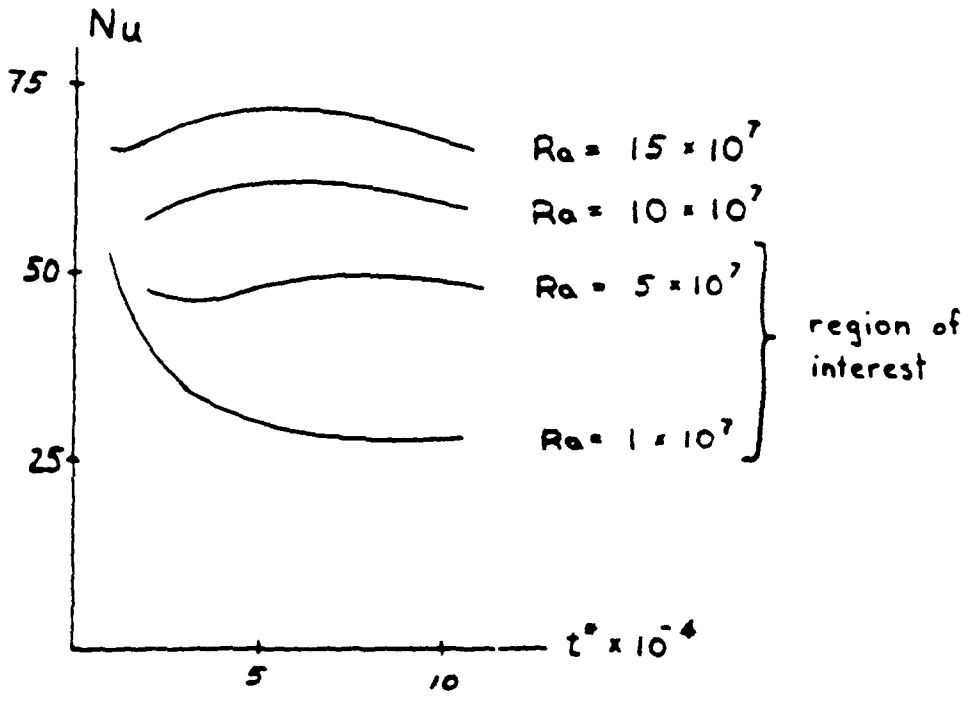
$$q = A\left(h_{rad} + h_{conv}\right)\left(T_s - T_f\right) \tag{5}$$

where

$$h_{rad} = \varepsilon\sigma\left(T_{avg} + T_f\right)\left(T_{avg}^2 + T_f^2\right) \tag{6}$$

$$h_{conv} = 1.42\left(\frac{\Delta T}{L}\right)^{1/4} \tag{7}$$

The convection heat transfer coefficient is calculated from an empirical formula for natural convection from a vertical plate, Ozisik (2). Transient natural convection from a vertical surface with an isothermal heated section was modeled numerically, and the time dependence of the mean Nusselt number for the heated region was calculated as a function of the Rayleigh number, Ra. The results of this model are shown in figure (3). Steady-state values of the Nusselt number, as predicted by empirical correlations in the literature, are within fifteen percent of the average transient values, for the range of Rayleigh numbers of interest. Natural convection losses from the front surface



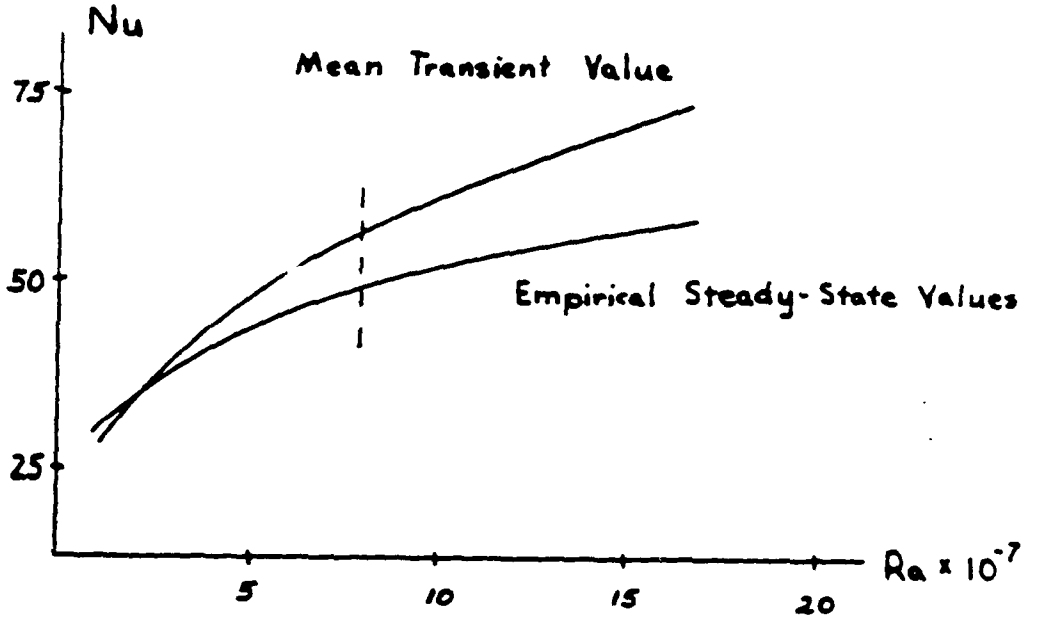


Figure 3

Transient Nusselt number for the front surface of the target plate and a comparison with the steady-state predictions of empirical correlations.

are very small in comparison with the absorbed laser energy, so the use of simplified steady-state correlations for the convection heat transfer coefficient is certainly justified.

Radiation and natural convection losses from the rear surface of the target plate are handled in exactly the same manner as on the front surface, except the emissivity of the rear surface is assumed to be .9. The high emissivity of the rear surface is due to the temperature sensitive paint.

The initial temperature of the target plate is 20°c. This is also the assumed environmental temperature. The target plate very quickly heats up to temperatures close to its melting point. Thermophysical properties vary considerably over this wide temperature range, and the temperature dependence of the thermal conductivity and the specific heat are included in the model. The target plate is assumed to be made from 304 stainless steel, with a thickness of 1.8 mm. The source of the thermophysical properties is Taylor, (3). Empirical correlations of the measured data are used. For the thermal conductivity of 304 stainless steel, the following relation applies:

$$k = .00015419\ T + .131 \quad (w/cm\ K) \qquad (8)$$

where the temperature is in degrees celsius. For the specific heat of stainless steel:

$$C_p = a_o + a_1 x + a_2 x^2 + a_3 x^3 \qquad (9)$$

where

$a_o = .4818888892$
$a_1 = .3225589198$
$a_2 = -.354040397$
$a_3 = .2037036995$
$x = T°c/1000$

The finite-difference equations are derived by application of equation (1), the conservation of energy principle, to the nodal control volumes. Application of this principle to the control volume shown in figure (2) results in the following relation for the interior nodes:

$$\rho C_p \Delta x \Delta y \Delta z \left(\frac{T_{ijk}^{n+1} - T_{ijk}^{n}}{\Delta t} \right) = -k_1 \Delta y \Delta z \left(\frac{T_{ijk}^{n+1} - T_{i-1,jk}^{n+1}}{\Delta x} \right)$$
$$+ k_2 \Delta y \Delta z \left(\frac{T_{i+1,jk}^{n+1} - T_{ijk}^{n+1}}{\Delta x} \right) - k_3 \Delta x \Delta z \left(\frac{T_{ijk}^{n} - T_{ij-1k}^{n}}{\Delta y} \right)$$
$$+ k_4 \Delta x \Delta z \left(\frac{T_{ij+1k}^{n} - T_{ijk}^{n}}{\Delta y} \right) - k_5 \Delta x \Delta y \frac{T_{ijk}^{n} - T_{ijk-1}^{n}}{\Delta z} \qquad (10)$$
$$+ k_6 \Delta x \Delta y \left(\frac{T_{ijk+1}^{n} - T_{ijk}^{n}}{\Delta z} \right)$$

The heat transfer rates in the x direction, through the plate thickness, are treated implicitly. The thermal conductivities are evaluated at the surface temperatures

of the control volume, and the specific heat is evaluated at the temperature of the center node. Equation (10) is manipulated to put the implicit terms on the left side and the explicit terms on the right side:

$$T_{ijk}^{n+1} + \frac{\alpha_1 \Delta t}{\Delta x^2}\left(T_{ijk}^{n+1} - T_{i-1jk}^{n+1}\right) - \frac{\alpha_2 \Delta t}{\Delta x^2}\left(T_{i+1jk}^{n+1} - T_{ijk}^{n+1}\right) = T_{ijk}^{n}$$

$$- \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T_{ijk}^{n} - T_{ij-1k}^{n}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T_{ij+1k}^{n} - T_{ijk}^{n}\right) \qquad (11)$$

$$- \frac{\alpha_5 \Delta t}{\Delta z^2}\left(T_{ijk}^{n} - T_{ijk-1}^{n}\right) + \frac{\alpha_6 \Delta t}{\Delta z^2}\left(T_{ijk+1}^{n} - T_{ijk}^{n}\right)$$

The finite difference equations for the surface nodes of the computational domain are derived in the same manner. Figure (4) is a schematic of the nodal control volumes for the front and rear surfaces of the target plate. The finite-difference equation for the front surface nodes is:

$$\left(1 + \frac{2\alpha_2 \Delta t}{\Delta x^2}\right) T_{1jk}^{n+1} - \frac{2\alpha_2 \Delta t}{\Delta x^2} T_{2jk}^{n+1} = \frac{2 q_{jk} \Delta t}{\rho C_p \Delta x} + T_{1jk}^{n}$$

$$+ \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T_{1j-1k}^{n} - T_{1jk}^{n}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T_{1j+1k}^{n} - T_{1jk}^{n}\right) \qquad (12)$$

$$+ \frac{\alpha_5 \Delta t}{\Delta z^2}\left(T_{1jk-1}^{n} - T_{1jk}^{n}\right) + \frac{\alpha_6 \Delta t}{\Delta z^2}\left(T_{1jk+1}^{n} - T_{1jk}^{n}\right)$$

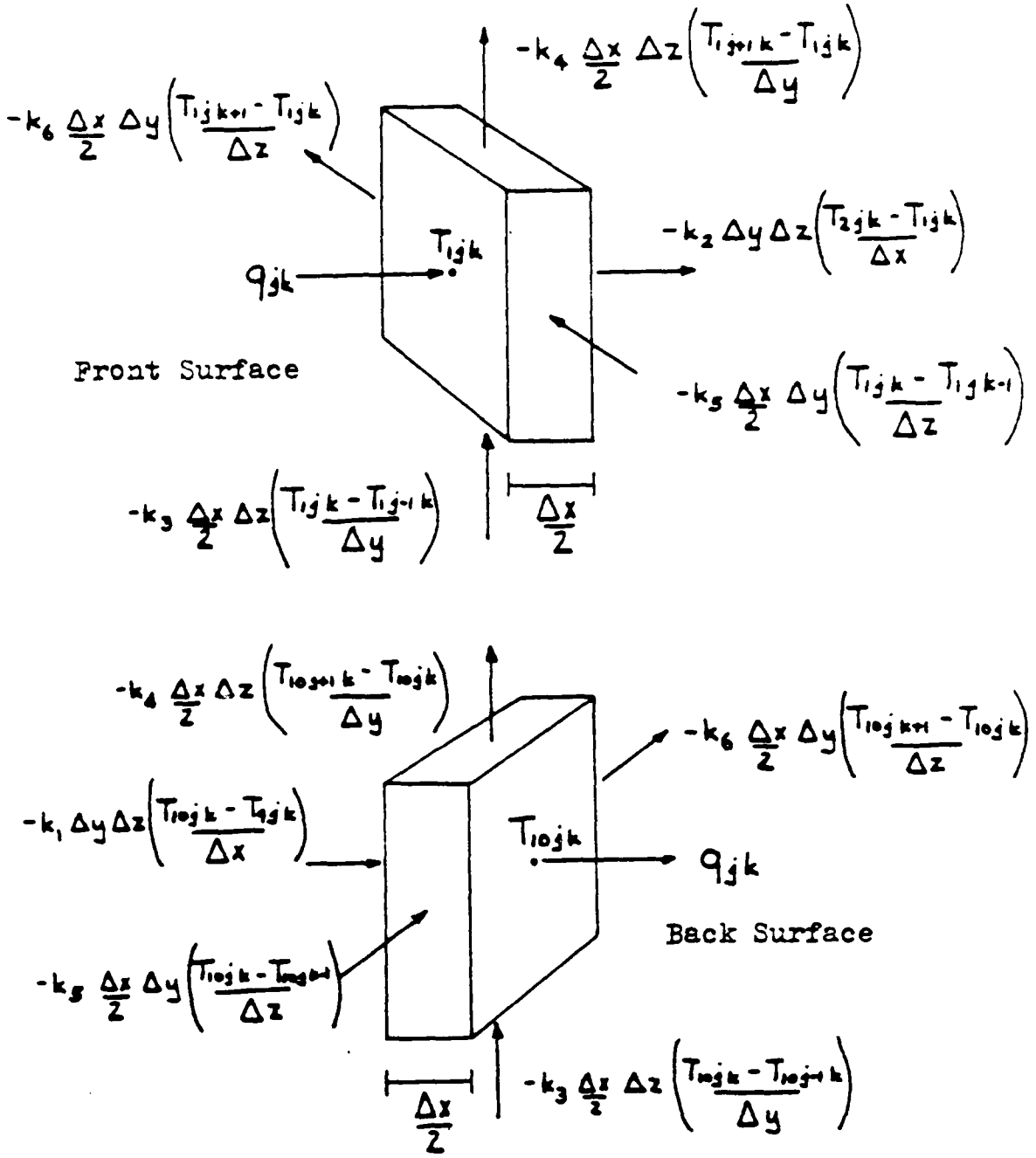


Figure 4

Schematic of the nodal control volumes, with the associated heat transfer rates crossing the control surface, for the front and rear surfaces of the target plate.

The finite-difference equations for the rear surface nodes are:

$$-2\frac{\alpha_1 \Delta t}{\Delta x^2} T_{9jk}^{n+1} + \left(1 + 2\frac{\alpha_1 \Delta t}{\Delta x^2}\right) T_{10jk}^{n+1} = -2\frac{q_{jk} \Delta t}{\rho C_p \Delta x} + T_{10jk}^{n}$$

$$+\frac{\alpha_3 \Delta t}{\Delta y^2}\left(T_{10j+1k}^{n} - T_{10jk}^{n}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T_{10j+1k}^{n} - T_{10jk}^{n}\right) \qquad (13)$$

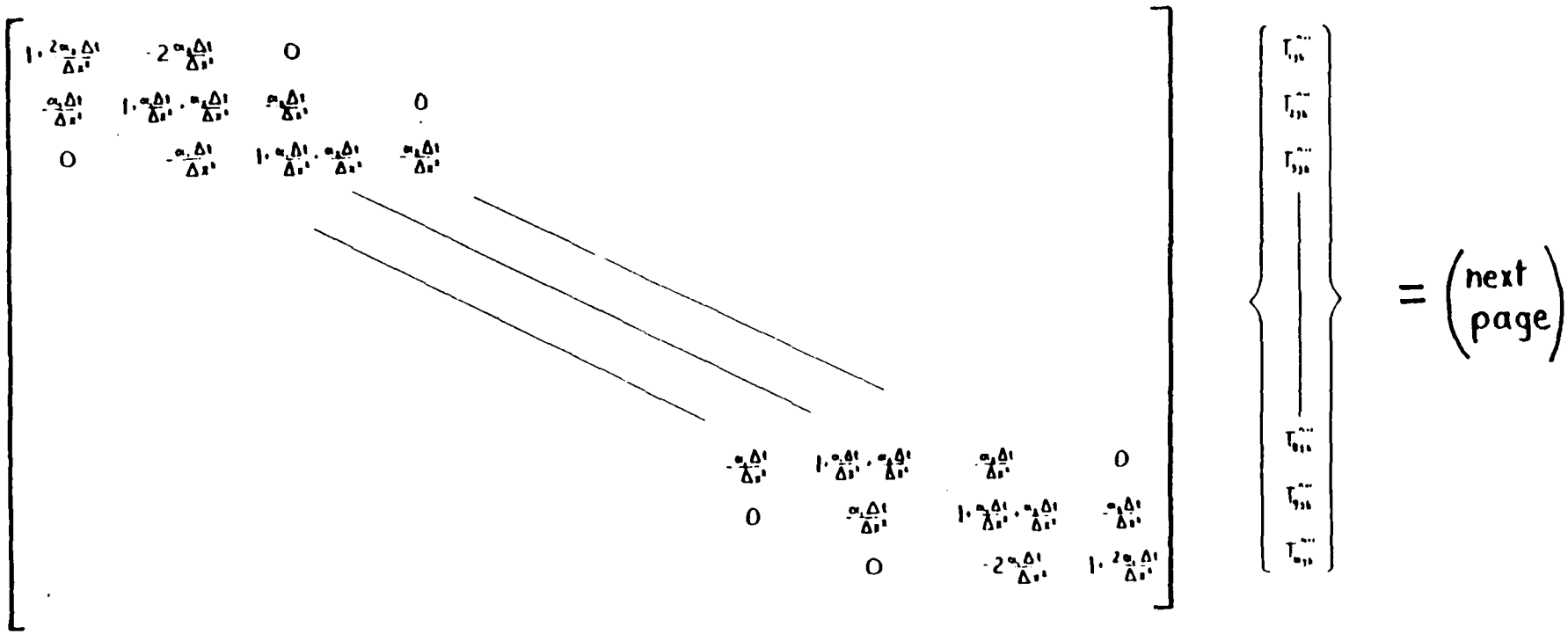
$$+\frac{\alpha_5 \Delta t}{\Delta z^2}\left(T_{10jk-1}^{n} - T_{10jk}^{n}\right) + \frac{\alpha_6 \Delta t}{\Delta z^2}\left(T_{10jk+1}^{n} - T_{10jk}^{n}\right)$$

The q term represents the net heat transfer rate through the surface, and it includes radiation and convection losses.

The ten nodal equations of a single row of node points in the x direction form a matrix equation, similar to equation (3). The matrix equation for the interior x direction rows is written out on the next two pages.

The four edge surfaces and the four corners of the target plate have similar governing matrix equations. The regions of applicability of these equations are shown in figure (5). The tri-diagonal coefficient matrix has the same form for all of these equations, so only the vectors on the right hand side of the equations are written out.

The ten simultaneous equations in a single matrix equation are solved using the Thomas algorithm, Roach (4). The solution technique is written out on page 31.



$$=$$

$$\left\{\begin{array}{l}
\frac{2q_w\Delta t}{\rho C_p \Delta x} + T_{1,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{1,j+1,k}^n - T_{1,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{1,j-1,k}^n - T_{1,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{1,j,k+1}^n - T_{1,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{1,j,k-1}^n - T_{1,j,k}^n\right) \\
T_{2,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{2,j+1,k}^n - T_{2,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{2,j-1,k}^n - T_{2,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{2,j,k+1}^n - T_{2,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{2,j,k-1}^n - T_{2,j,k}^n\right) \\
T_{3,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{3,j+1,k}^n - T_{3,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{3,j-1,k}^n - T_{3,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{3,j,k+1}^n - T_{3,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{3,j,k-1}^n - T_{3,j,k}^n\right) \\
\vdots \\
T_{i,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{i,j+1,k}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{i,j-1,k}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{i,j,k+1}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{i,j,k-1}^n - T_{i,j,k}^n\right) \\
T_{i,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{i,j+1,k}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{i,j-1,k}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{i,j,k+1}^n - T_{i,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{i,j,k-1}^n - T_{i,j,k}^n\right) \\
\frac{2q_w\Delta t}{\rho C_p \Delta x} + T_{m,j,k}^n + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{m,j+1,k}^n - T_{m,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta y^2}\left(T_{m,j-1,k}^n - T_{m,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{m,j,k+1}^n - T_{m,j,k}^n\right) + \frac{\alpha_2\Delta t}{\Delta z^2}\left(T_{m,j,k-1}^n - T_{m,j,k}^n\right)
\end{array}\right\}$$

Tri-diagonal matrix equation for the interior nodes of the computational domain.

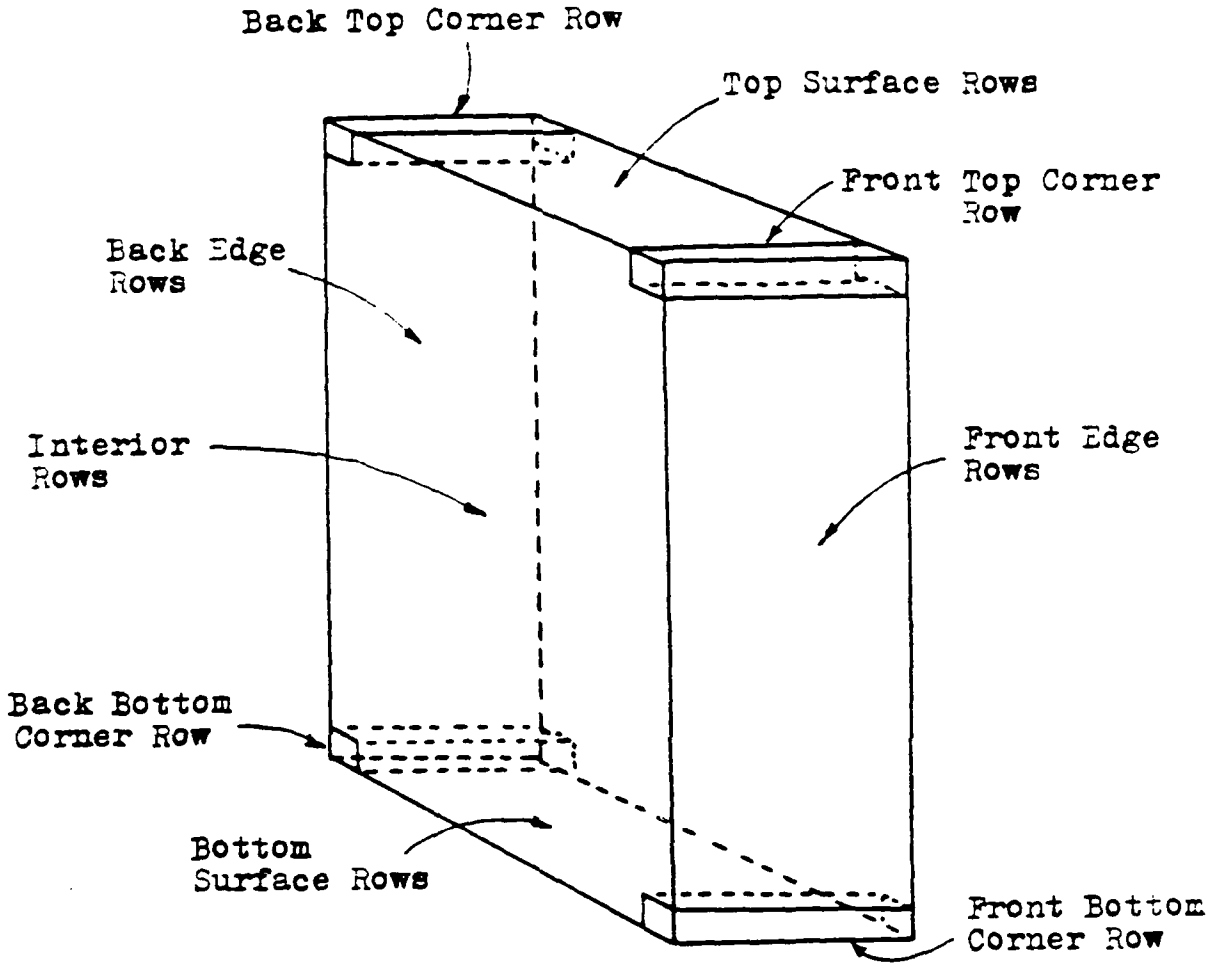


Figure 5

Schematic of the computational domain, showing the boundary rows of nodes that are treated implicitly.

$$
\left\{
\begin{array}{r}
\frac{2q_{11}\Delta t}{\rho c_p \Delta x} + T^n_{1,1} + \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T^n_{1,m-1} - T^n_{1,1}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{1,m-1} - T^n_{1,1}\right) + \frac{2\alpha_5 \Delta t}{\Delta z^2}\left(T^n_{1,2} - T^n_{1,1}\right) \\
T^n_{2,1} + \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T^n_{2,m-1} - T^n_{2,1}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{2,m-1} - T^n_{2,1}\right) + \frac{2\alpha_5 \Delta t}{\Delta z^2}\left(T^n_{2,2} - T^n_{2,1}\right) \\
\vdots \\
T^n_{q,1} + \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T^n_{q,m-1} - T^n_{q,1}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{q,m-1} - T^n_{q,1}\right) + \frac{2\alpha_5 \Delta t}{\Delta z^2}\left(T^n_{q,2} - T^n_{q,1}\right) \\
\frac{2q_{21}\Delta t}{\rho c_p \Delta x} + T^n_{m,1} + \frac{\alpha_3 \Delta t}{\Delta y^2}\left(T^n_{m,m-1} - T^n_{m,1}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{m,m-1} - T^n_{m,1}\right) + \frac{2\alpha_5 \Delta t}{\Delta z^2}\left(T^n_{m,2} - T^n_{m,1}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the front edge rows.

$$
\left\{
\begin{array}{r}
\dfrac{2q_{m,n}\Delta t}{\rho C_{p_{m,n}}\Delta x} + T^{n}_{1,m,n} + \dfrac{2\alpha_3\Delta t}{\Delta y^2}\left(T^{n}_{1,m+1,n} - T^{n}_{1,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{1,m,n+1} - T^{n}_{1,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{1,m,n-1} - T^{n}_{1,m,n}\right) \\
T^{n}_{2,m,n} + \dfrac{2\alpha_3\Delta t}{\Delta y^2}\left(T^{n}_{2,m+1,n} - T^{n}_{2,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{2,m,n+1} - T^{n}_{2,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{2,m,n-1} - T^{n}_{2,m,n}\right) \\
\vdots \\
T^{n}_{q,m,n} + \dfrac{2\alpha_3\Delta t}{\Delta y^2}\left(T^{n}_{q,m+1,n} - T^{n}_{q,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{q,m,n+1} - T^{n}_{q,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{q,m,n-1} - T^{n}_{q,m,n}\right) \\
\dfrac{2q_{m,n}\Delta t}{\rho C_{p_{m,n}}\Delta x} + T^{n}_{m,m,n} + \dfrac{2\alpha_3\Delta t}{\Delta y^2}\left(T^{n}_{m,m+1,n} - T^{n}_{m,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{m,m,n+1} - T^{n}_{m,m,n}\right) + \dfrac{\alpha_3\Delta t}{\Delta z^2}\left(T^{n}_{m,m,n-1} - T^{n}_{m,m,n}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the top surface rows.

$$
\left\{
\begin{array}{r}
\frac{2q_{1,jm}\Delta t}{\rho C_p \Delta_3} + T^n_{1,jm} + \frac{\alpha_1 \Delta t}{\Delta y^2}\left(T^n_{1,jm-1} - T^n_{1,jm}\right) + \frac{\alpha_0 \Delta t}{\Delta y^2}\left(T^n_{1,jm+1} - T^n_{1,jm}\right) + \frac{2\alpha_2 \Delta t}{\Delta z^2}\left(T^n_{1,j,m} - T^n_{1,jm}\right) \\
T^n_{2,jm} + \frac{\alpha_1 \Delta t}{\Delta y^2}\left(T^n_{2,jm-1} - T^n_{2,jm}\right) + \frac{\alpha_2 \Delta t}{\Delta y^2}\left(T^n_{2,jm+1} - T^n_{2,jm}\right) + \frac{2\alpha_2 \Delta t}{\Delta z^2}\left(T^n_{2,j,m} - T^n_{2,jm}\right) \\
\vdots \\
T^n_{q,jm} + \frac{\alpha_1 \Delta t}{\Delta y^2}\left(T^n_{q,jm-1} - T^n_{q,jm}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{q,jm+1} - T^n_{q,jm}\right) + \frac{2\alpha_2 \Delta t}{\Delta z^2}\left(T^n_{q,j,m} - T^n_{q,jm}\right) \\
\frac{2q_{m,jm}\Delta t}{\rho C_p \Delta_3} + T^n_{m,jm} + \frac{\alpha_1 \Delta t}{\Delta y^2}\left(T^n_{m,jm-1} - T^n_{m,jm}\right) + \frac{\alpha_4 \Delta t}{\Delta y^2}\left(T^n_{m,jm+1} - T^n_{m,jm}\right) + \frac{2\alpha_2 \Delta t}{\Delta z^2}\left(T^n_{m,j,m} - T^n_{m,jm}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the back edge rows.

$$
\left\{
\begin{array}{r}
\frac{2q_{1R}\Delta t}{\rho C_p \Delta z} + T^n_{11R} + \frac{2\alpha_4\Delta t}{\Delta y^2}\left(T^n_{12R} - T^n_{11R}\right) + \frac{\alpha_3\Delta t}{\Delta z^2}\left(T^n_{11RW} - T^n_{11R}\right) + \frac{\alpha_5\Delta t}{\Delta z^2}\left(T^n_{11RW} - T^n_{11R}\right) \\
T^n_{21R} + \frac{2\alpha_4\Delta t}{\Delta y^2}\left(T^n_{22R} - T^n_{21R}\right) + \frac{\alpha_3\Delta t}{\Delta z^2}\left(T^n_{21RW} - T^n_{21R}\right) + \frac{\alpha_5\Delta t}{\Delta z^2}\left(T^n_{21RW} - T^n_{21R}\right) \\
\vdots \\
T^n_{91R} + \frac{2\alpha_4\Delta t}{\Delta y^2}\left(T^n_{92R} - T^n_{91R}\right) + \frac{\alpha_3\Delta t}{\Delta z^2}\left(T^n_{91RW} - T^n_{91R}\right) + \frac{\alpha_5\Delta t}{\Delta z^2}\left(T^n_{91RW} - T^n_{91R}\right) \\
\frac{2q_{1R}\Delta t}{\rho C_p \Delta z} + T^n_{101R} + \frac{2\alpha_4\Delta t}{\Delta y^2}\left(T^n_{102R} - T^n_{101R}\right) + \frac{\alpha_3\Delta t}{\Delta z^2}\left(T^n_{101RW} - T^n_{101R}\right) + \frac{\alpha_5\Delta t}{\Delta z^2}\left(T^n_{101RW} - T^n_{101R}\right)
\end{array}
\right\}
$$

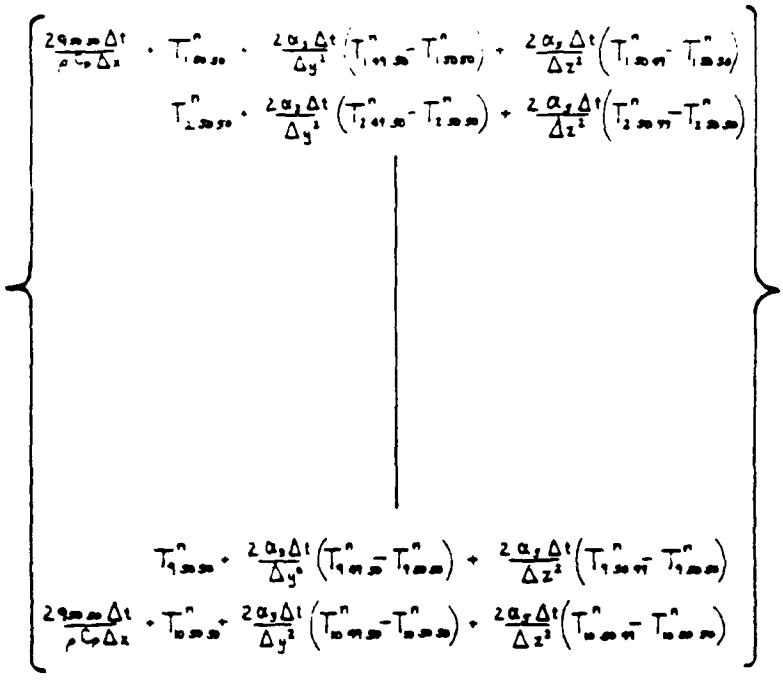
Right hand side of the matrix equation for the bottom surface rows.

$$
\left\{
\begin{array}{l}
\dfrac{2q_{m,1}\Delta t}{\rho C_p \Delta z} + T^n_{1,n,1} + \dfrac{2\alpha_2\Delta t}{\Delta y^2}\left(T^n_{1,n+1,1} - T^n_{1,n,1}\right) + \dfrac{2\alpha_3\Delta t}{\Delta z^2}\left(T^n_{1,n,2} - T^n_{1,n,1}\right) \\
T^n_{2,n,1} + \dfrac{2\alpha_2\Delta t}{\Delta y^2}\left(T^n_{2,n+1,1} - T^n_{2,n,1}\right) + \dfrac{2\alpha_3\Delta t}{\Delta z^2}\left(T^n_{2,n,2} - T^n_{2,n,1}\right) \\
\vdots \\
T^n_{q,n,1} + \dfrac{2\alpha_2\Delta t}{\Delta y^2}\left(T^n_{q,n+1,1} - T^n_{q,n,1}\right) + \dfrac{2\alpha_3\Delta t}{\Delta z^2}\left(T^n_{q,n,2} - T^n_{q,n,1}\right) \\
\dfrac{2q_{m,1}\Delta t}{\rho C_p \Delta z} + T^n_{m,n,1} + \dfrac{2\alpha_2\Delta t}{\Delta y^2}\left(T^n_{m,n+1,1} - T^n_{m,n,1}\right) + \dfrac{2\alpha_3\Delta t}{\Delta z^2}\left(T^n_{m,n,2} - T^n_{m,n,1}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the front top corner row.

$$
\left\{
\begin{array}{l}
\dfrac{2q_0\Delta t}{\rho C_p \Delta z} + T^n_{111} + \dfrac{2\alpha_1\Delta t}{\Delta y^2}\left(T^n_{121} - T^n_{111}\right) + \dfrac{2\alpha_2\Delta t}{\Delta z^2}\left(T^n_{112} - T^n_{111}\right) \\
\qquad T^n_{211} + \dfrac{2\alpha_1\Delta t}{\Delta y^2}\left(T^n_{221} - T^n_{211}\right) + \dfrac{2\alpha_1\Delta t}{\Delta z^2}\left(T^n_{212} - T^n_{211}\right) \\
\qquad\qquad \vdots \\
\qquad T^n_{911} + \dfrac{2\alpha_1\Delta t}{\Delta y^2}\left(T^n_{921} - T^n_{911}\right) + \dfrac{2\alpha_1\Delta t}{\Delta z^2}\left(T^n_{912} - T^n_{911}\right) \\
\dfrac{2q_0\Delta t}{\rho C_p \Delta z} + T^n_{m11} + \dfrac{2\alpha_0\Delta t}{\Delta y^2}\left(T^n_{m21} - T^n_{m11}\right) + \dfrac{2\alpha_0\Delta t}{\Delta z^2}\left(T^n_{m12} - T^n_{m11}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the front bottom corner row.



Right hand side of the matrix equation for the back top corner row.

$$
\left\{
\begin{array}{l}
\dfrac{2q_{1,m}\Delta t}{\rho c_p \Delta z} + T^n_{1,m} + \dfrac{2\alpha_x \Delta t}{\Delta y^2}\left(T^n_{1,m-1} - T^n_{1,m}\right) + \dfrac{2\alpha_z \Delta t}{\Delta z^2}\left(T^n_{1,m} - T^n_{1,m}\right) \\
\qquad T^n_{2,m} + \dfrac{2\alpha_x \Delta t}{\Delta y^2}\left(T^n_{2,m-1} - T^n_{2,m}\right) + \dfrac{2\alpha_z \Delta t}{\Delta z^2}\left(T^n_{2,m-1} - T^n_{2,m}\right) \\
\qquad\qquad \vdots \\
\qquad T^n_{q,m} + \dfrac{2\alpha_x \Delta t}{\Delta y^2}\left(T^n_{q,m-1} - T^n_{q,m}\right) + \dfrac{2\alpha_z \Delta t}{\Delta z^2}\left(T^n_{q,m-1} - T^n_{q,m}\right) \\
\dfrac{2q_{n,m}\Delta t}{\rho c_p \Delta z} + T^n_{n,m} + \dfrac{2\alpha_y \Delta t}{\Delta y^2}\left(T^n_{n,m-1} - T^n_{n,m}\right) + \dfrac{2\alpha_z \Delta t}{\Delta z^2}\left(T^n_{n,m-1} - T^n_{n,m}\right)
\end{array}
\right\}
$$

Right hand side of the matrix equation for the back bottom corner row.

Direct Solution of a Tri-diagonal System of Equations by Gaussian Elimination:

$$
\begin{aligned}
b_1u_1 + c_1u_2 &= d_1 \\
a_2u_1 + b_2u_2 + c_2u_3 &= d_2 \\
a_3u_2 + b_3u_3 + c_3u_4 &= d_3 \\
&\cdots\cdots\cdots\cdots \\
a_{n-1}u_{n-2} + b_{n-1}u_{n-1} + c_{n-1}u_n &= d_{n-1} \\
a_nu_{n-1} + b_nu_n &= d_n
\end{aligned}
$$

Use equation 1 to eliminate u_1 from equation 2. Use this new equation 2 to eliminate u_2 from equation 3. Continue to end. Last equation will then yield an explicit expression for u_n:

$$u_n = \gamma_n \ .$$

Using this value of u_n, substitute back into equation n and solve for u_{n-1}. Continue back-substitution. u_i given by:

$$u_i = \gamma_i - \frac{c_iu_{i+1}}{\beta_i} \quad (i = n-1,\ n-2,\ \ldots.,\ 1)$$

The coefficients β and γ can be calculated in advance from

$$\beta_1 = b_1, \qquad \gamma_1 = \frac{d_1}{\beta_1}$$

$$\beta_i = b_i - \frac{a_ic_{i-1}}{\beta_{i-1}}, \qquad \gamma_i = \frac{d_i - a_i\gamma_{i-1}}{\beta_i} \quad (i = 2,3,\ldots.,n)$$

Thus, once the β's and γ's are determined, the u_i can be calculated directly starting with i=n and going to i=1.

The advantage of an implicit numerical scheme over an explicit scheme lies in the restrictive timestep limitations of explicit schemes. The stability of implicit equations does not restrict the maximum size of a timestep. The maximum size of a timestep is restricted by the numerical stability of explicit equations. The maximum allowable timestep sizes of explicit equations are functions of the mesh spacing and the thermal diffusivity of the material. Decreasing the mesh size and increasing the dimensionality of the numerical model can make the timestep size limitation very restrictive and costly in computational time. Because of the large thermal gradients in the direction through the thickness of the plate, the distance between nodes is significantly shorter than in the other two directions, and this direction most severely restricts the maximum allowable timestep size for explicit finite-difference equations. The beneficial effect of treating one direction implicitly is therefore greatest in this direction.

The stability limitation on the timestep size for explicit finite-difference equations can be determined by putting the equations in the following form:

$$\phi_{ij}^{n+1} = a_1 \phi_{ij}^{n} + a_2 \phi_{i+1,j}^{n} + a_3 \phi_{i-1,j}^{n} + a_4 \phi_{i,j+1}^{n} + a_5 \phi_{i,j-1}^{n} \qquad (14)$$

As long as the coefficients of the explicit terms are positive, the equation is stable. The various finite-difference

equations can have different timestep restrictions, due to the variation in mesh size and thermophysical properties of the material. Convection and radiation boundary conditions can also adversely effect the maximum allowable timestep size. The most restrictive timestep criteria for the computational mesh should be used.

The transient solutions for implicit and explicit equations can differ, if too large an implicit timestep is used. Transient solutions are more accurately modeled using smaller timesteps. This is generally not a problem with explicit solutions. The two types of solutions approach each other as the number of iterations increase, and trial comparisons of the results of a completely explicit three-dimensional conduction model and a partially implicit model demonstrated that differences between the two types of solutions were negligible when the implicit solution had at least a hundred iterations.

The numerical model was run on a VAX 11/780 computer. Inputs to the model are the spatial distribution of the intensity of laser radiation, incident on the front surface of the target plate, and the exposure time. The output of the numerical model is the temperature distribution on the back surface of the target plate. Two models were run; one with three-dimensional conduction in the target plate, and the other that allowed conduction only in the direction normal

to the front surface of the target plate. Results are presented in the next section.

RESULTS

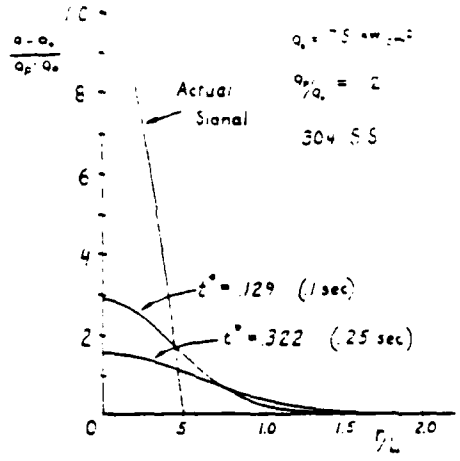
The target plate laser intensity measurement technique assumes that the conduction of heat through the target plate is one-dimensional. Errors introduced by this assumption will be greatest in regions with large thermal gradients in directions parallel to the plate surfaces, such as occur where there are narrow intensity spikes in the laser beam. To quantify the errors in the measurement, introduced by assuming one-dimensional heat transfer, one-dimensional and three-dimensional numerical solutions of problems with the same boundary conditions are compared.

If the problem of thermal conduction through the target plate is appropriately non-dimensionalized, there are only two independent parameters to the problem; the intensity distribution in the laser beam cross-section, and the elapsed time that the target plate is exposed to the laser beam before the measurement data is collected. The important parameter in the shape of intensity spikes is the spike width, normalized by the thickness of the target plate. The height of an isolated intensity spike does not influence the capability of the target plate laser intensity measurement technique to resolve it. For a pulse with a given spike width, the measurement technique will predict a spike with a maximum intensity that is some

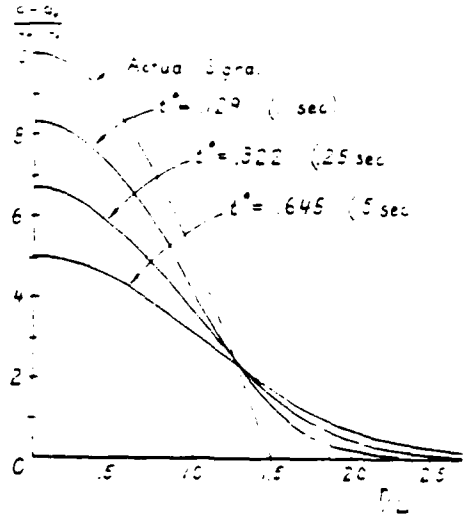
fraction of the actual peak intensity of the spike. The predicted fractional peak intensity of the spike is independent of the actual peak intensity of the incident spike, above the background laser intensity field. This was demonstrated in Shadday (5) and (6). This study looked at the capability of the laser intensity measurement technique to resolve a single axisymmetric intensity spike, superimposed on a uniform intensity background. Some results of this study are shown in figure (6). The predicted response to spikes with widths of one through four plate thicknesses are shown for several elapsed exposure times. The spike height above the background laser radiation is normalized by the peak height of the actual spike above the background intensity. The elapsed time is non-dimensionalized by the ratio of the square of the plate thickness to the reference thermal diffusivity:

$$t^* = \frac{\alpha_r t}{L^2} \qquad (15)$$

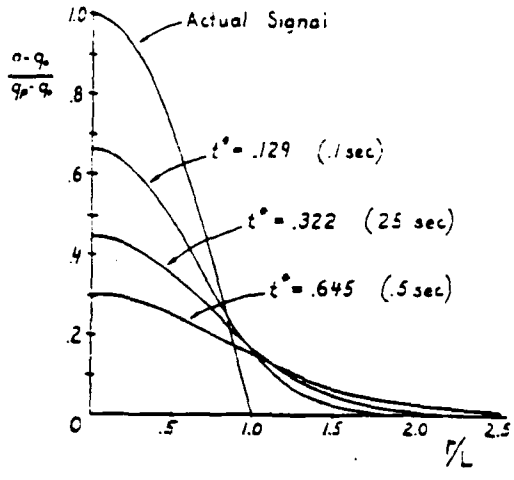
The non-dimensionalized time correlates the response of target plates made of different metals. Narrow intensity spikes, with widths less than four plate thicknesses, are resolved poorly by the laser intensity target plate measurement technique. As the exposure time of the target plate to the laser beam increases, the resolution of the measurement technique decreases.



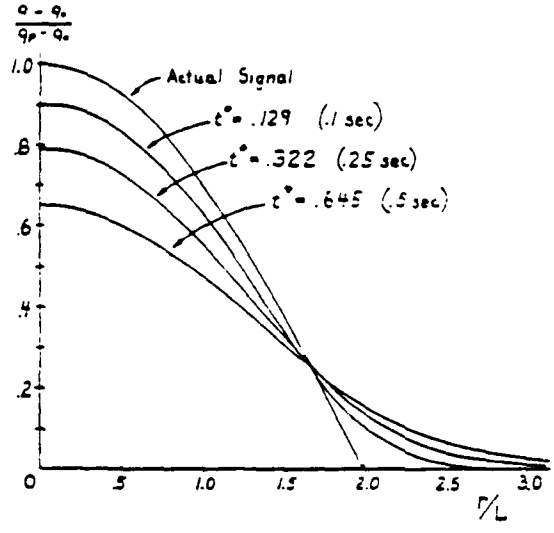
Spike width is one plate thickness



Spike width is three plate thicknesses



Spike width is two plate thicknesses



Spike width is four plate thicknesses

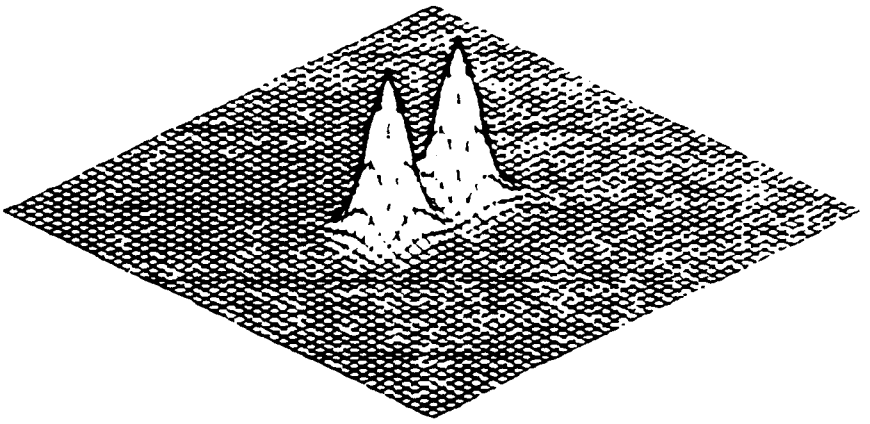
Figure 6

Predicted axisymmetric intensity spikes on the front surface of a target plate. From Shadday (5) and (6).

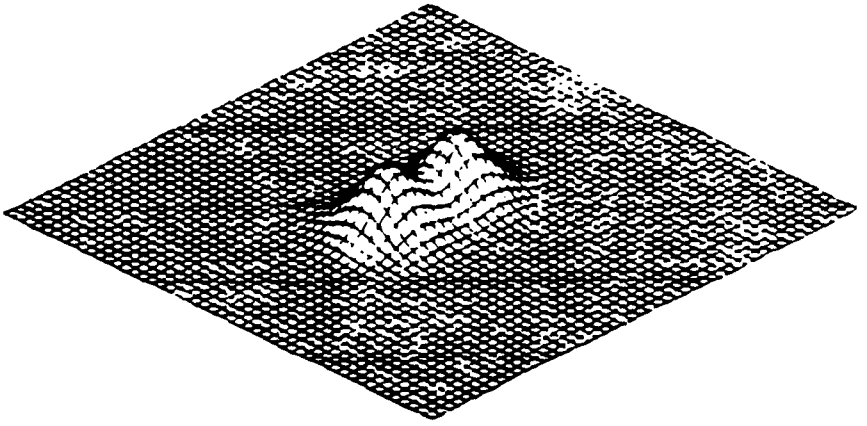
All of the cases run on the numerical model were for a target plate made of 304 stainless steel, with a thickness of 1.8 mm. Predicted laser beam intensity distributions are presented along with the actual intensity distribution for several cases. The intensity distributions are actually the temperature distributions on the rear surface of the target plate. If conduction through the plate is actually one-dimensional, the temperature distribution on the rear surface of the target plate will have exactly the same shape as the intensity distribution, incident upon the front surface. Therefore the target plate laser intensity measurement technique will predict an intensity distribution with the shape of the temperature distribution on the rear surface of the target plate. The difference between the intensity and the background intensity is normalized by the difference between the actual peak intensity and the background intensity:

$$q^* = \frac{q - q_0}{q_p - q_0} \tag{16}$$

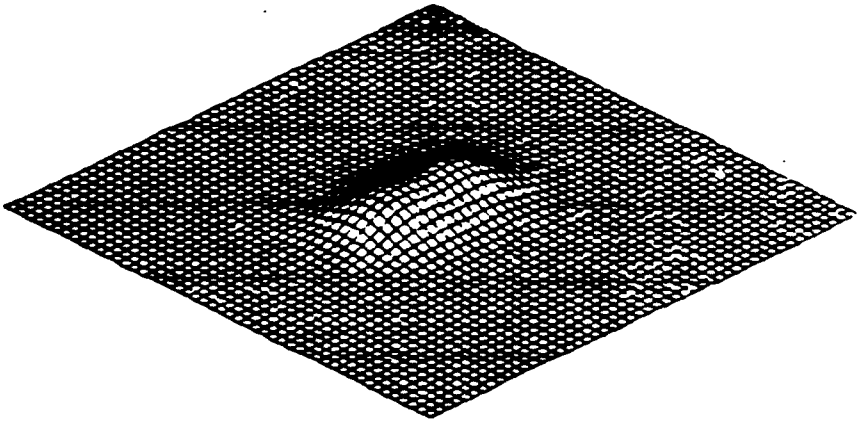
Figure (7) shows the predicted intensity distributions for two adjacent spikes, with widths of two plate thicknesses. The non-dimensional exposure times are .108 and .325. For the stainless steel target plate, these correspond to exposure times of .1 and .3 seconds respectively. Figure (8) shows predicted intensity profiles for the same two cases.



Actual intensity distribution



Predicted intensity distribution, $t^* = .108$



Predicted intensity distribution, $t^* = .325$

Figure 7

Predicted laser beam intensity distributions with spike widths of two plate thicknesses.

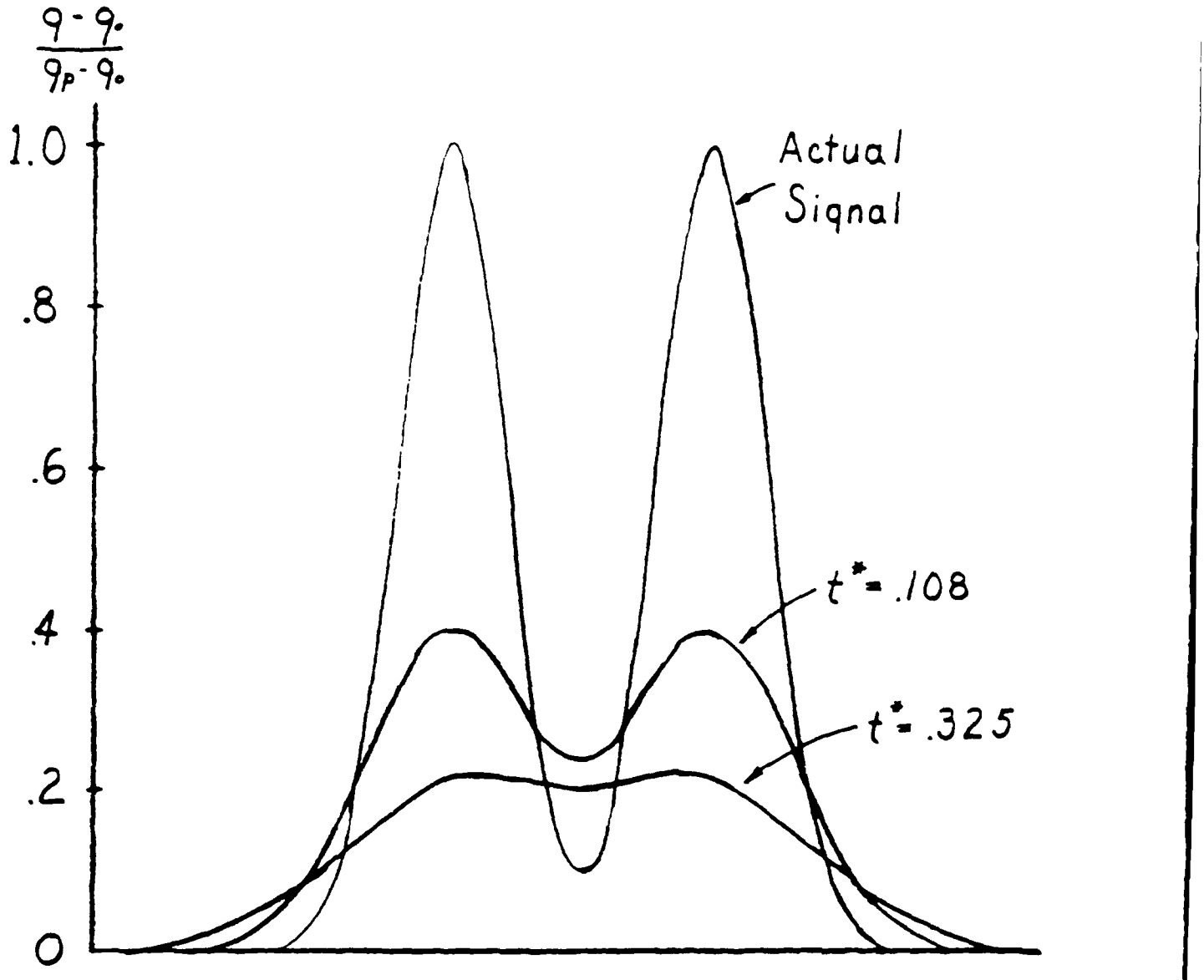


Figure 8

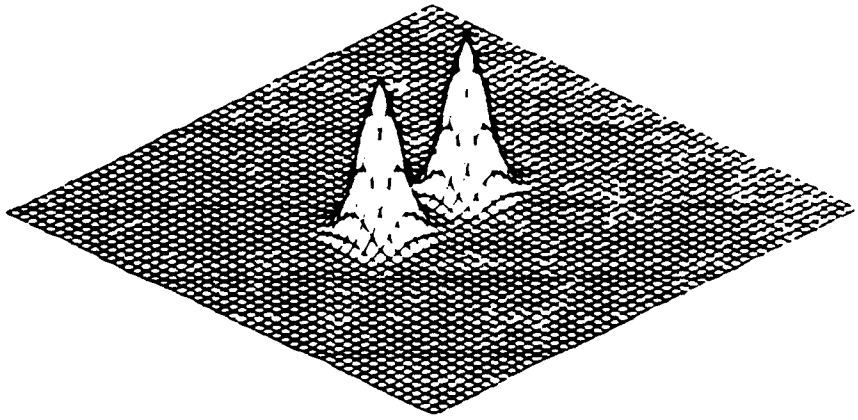
Predicted intensity profiles of spikes with a width of two plate thicknesses.

Figures (9) through (16) show the predicted intensity distributions for two adjacent spikes, with widths of three to six plate thicknesses. Figure (17) shows predicted intensity profiles for the same cases. The non-dimensional exposure times are .108, .325, and .542. The last exposure time corresponds to .5 seconds, for stainless steel.

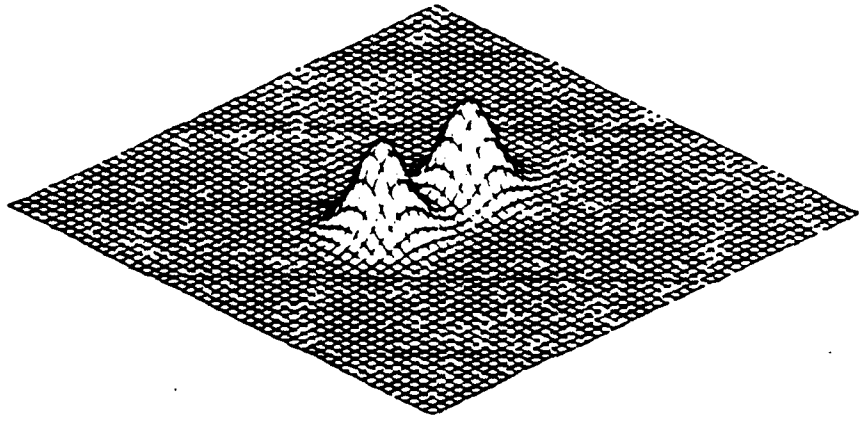
Figure (18) shows the predicted intensity distributions for two spikes with widths of four plate thicknesses. One spike has a peak intensity twice that of the other. The non-dimensional exposure times are .108 and .542. Figure (19) shows predicted intensity distributions for two overlapping spikes, with widths of four plate thicknesses. The distance between the two peaks is four plate thicknesses, and one spike has twice the peak intensity of the other. Figure (20) shows predicted intensity profiles for this input.

Figures (21) and (22) show predicted intensity distributions for a central spike, with a width of three plate thicknesses, surrounded by four spikes, with widths of six plate thicknesses. The central spike has twice the peak intensity of the four surrounding spikes.

Figure (23) shows the predicted intensity distributions for a laser beam, with a square cross-section 20 cm. on a side. In one corner, there are two intensity spikes with peak intensities of 28,000 and 20,000 w/cm^2 respectively. The background intensity is 6000 w/cm^2. The target plate



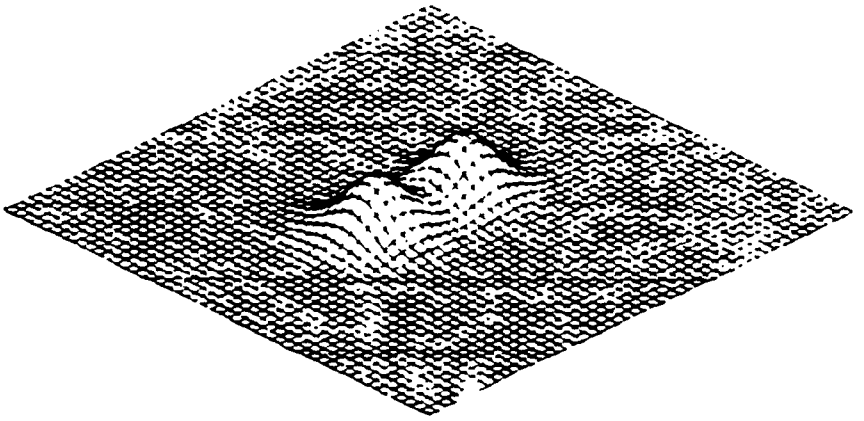
Actual intensity distribution



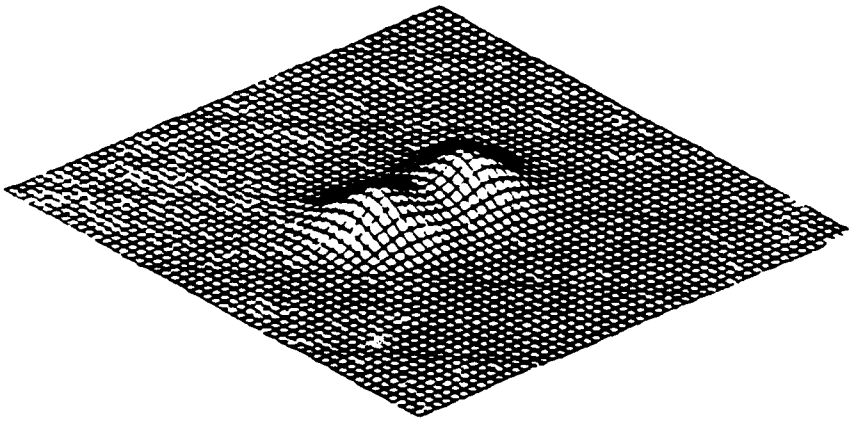
Predicted intensity distribution, t^{*}=.108

Figure 9

Predicted laser intensity distributions with spike widths of three plate thicknesses.



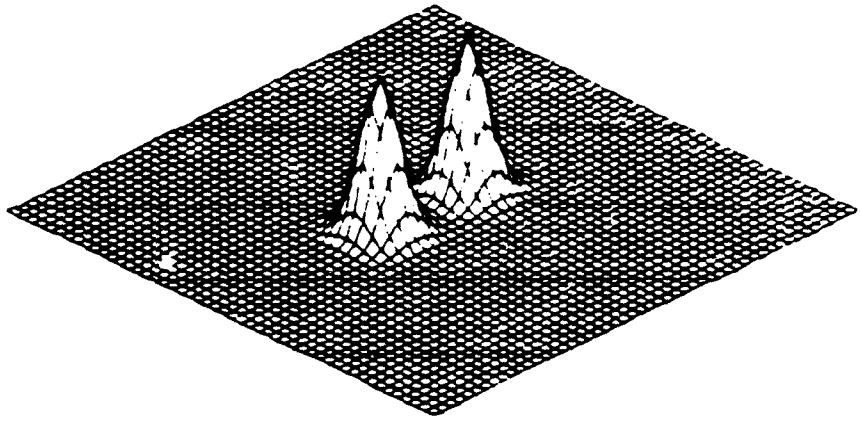
Predicted intensity distribution, $t^* = .325$



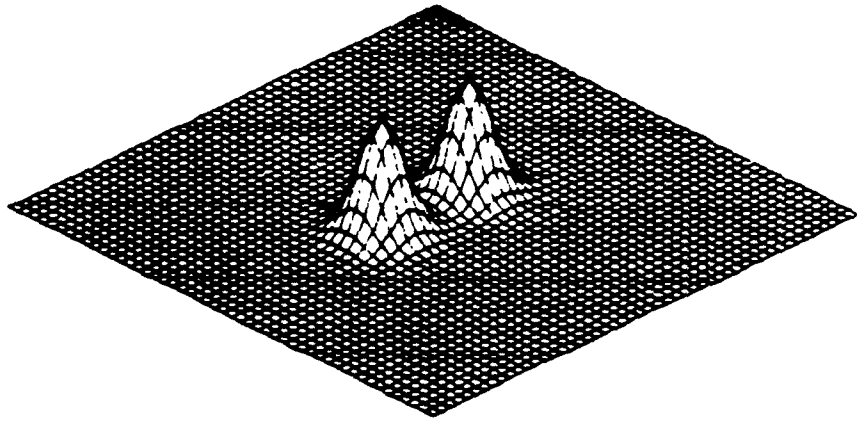
Predicted intensity distribution, $t^* = .542$

Figure 10

Predicted laser intensity distributions with spike widths of three plate thicknesses.



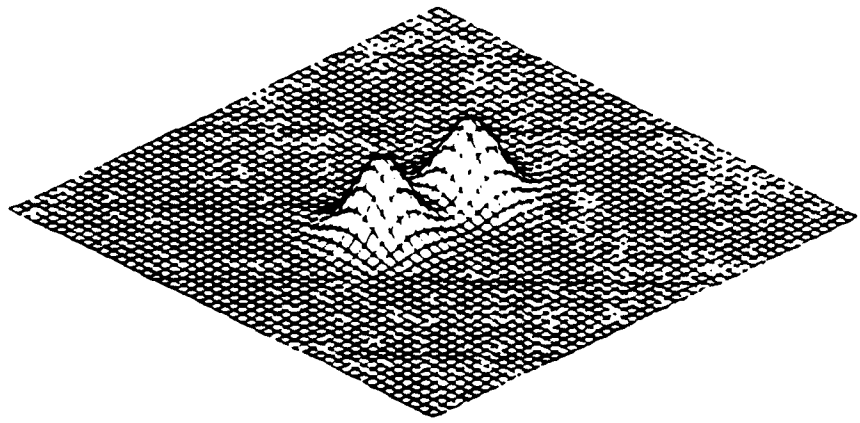
Actual intensity distribution



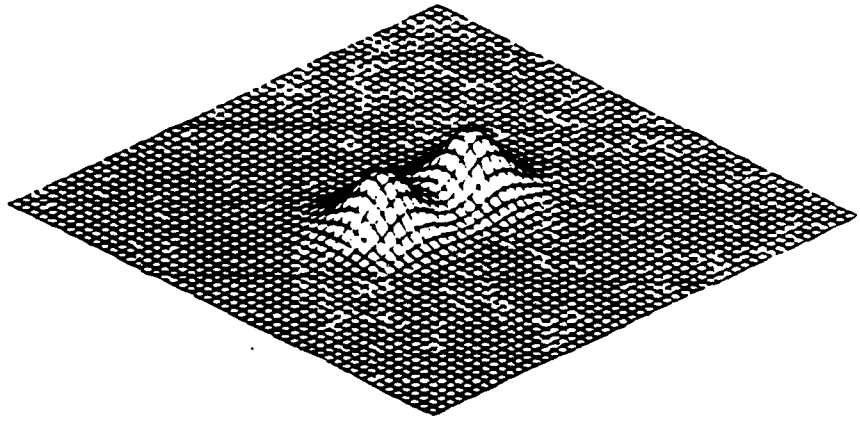
Predicted intensity distribution, $t^* = .108$

Figure 11

Predicted laser intensity distributions with spike widths of four plate thicknesses.



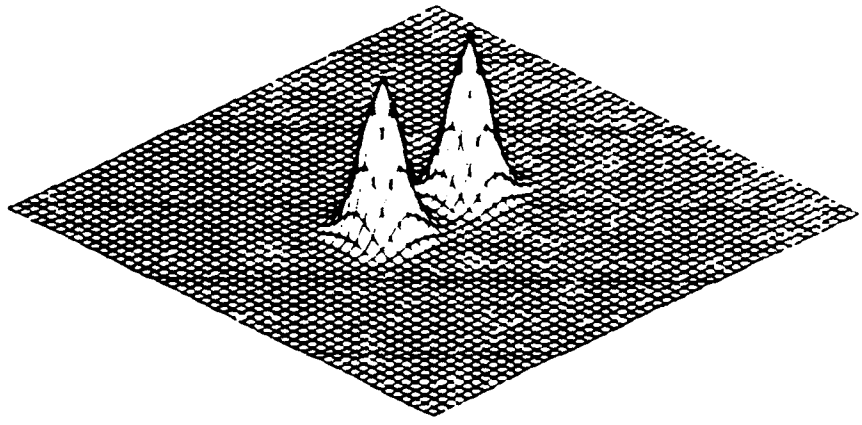
Predicted intensity distribution, $t^* = .325$



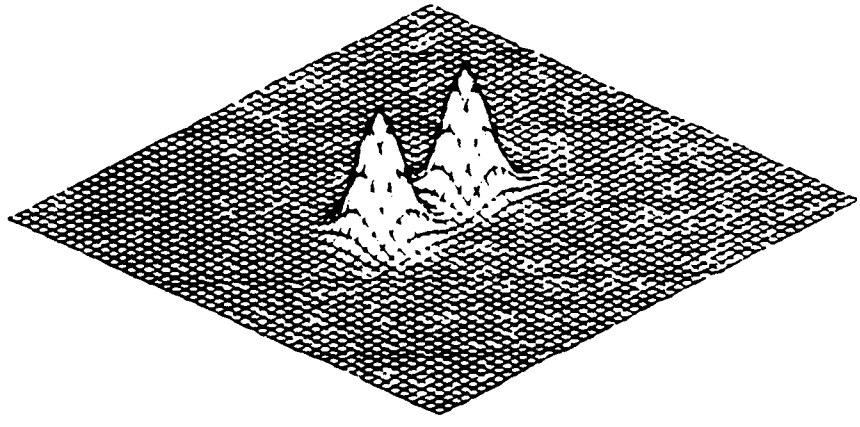
Predicted intensity distribution, $t^* = .542$

Figure 12

Predicted laser intensity distributions with spike widths of four plate thicknesses.



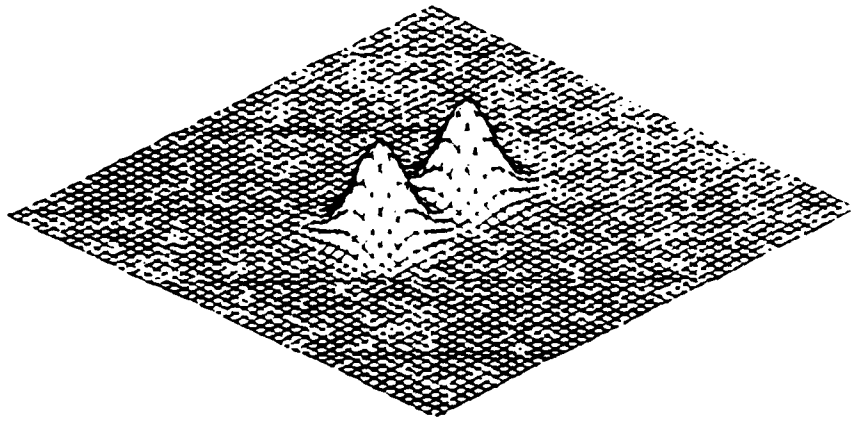
Actual intensity distribution



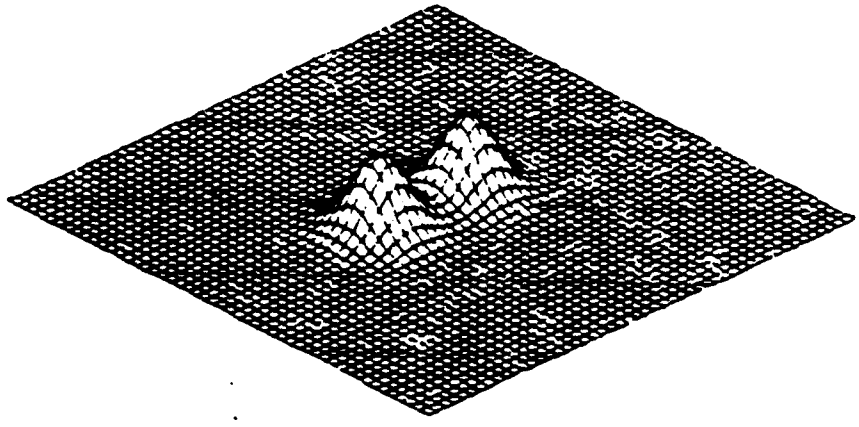
Predicted intensity distribution, $t^{*}=.108$

Figure 13

Predicted laser intensity distributions with spike widths of five olate thicknesses.



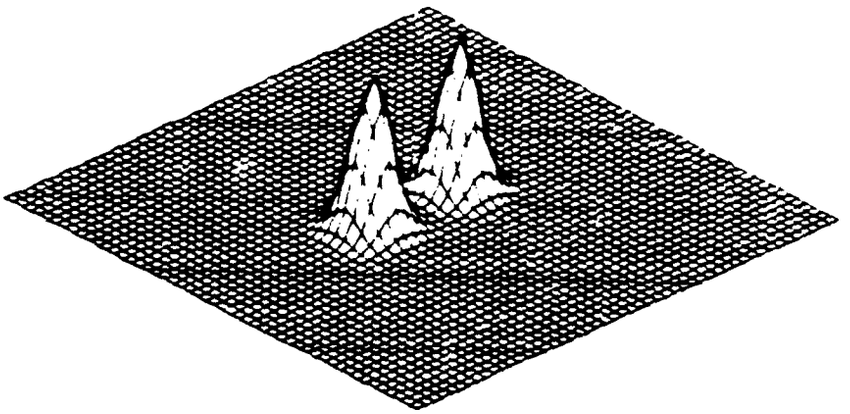
Predicted intensity distribution, $t^* = .325$



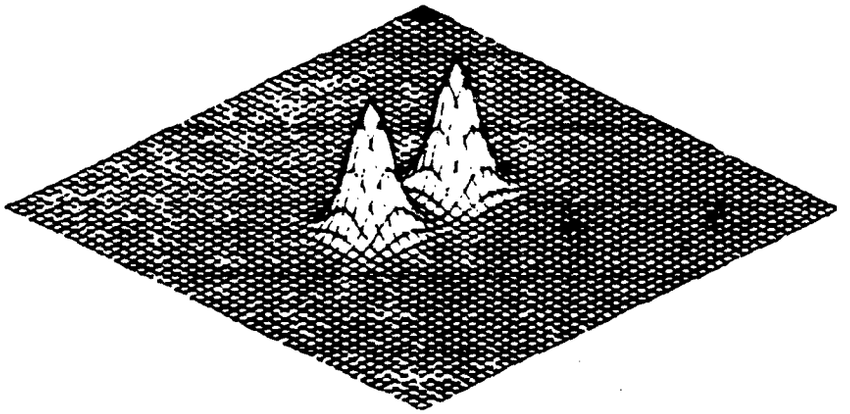
Predicted intensity distribution, $t^* = .542$

Figure 14

Predicted laser intensity distributions with spike widths of five plate thicknesses.



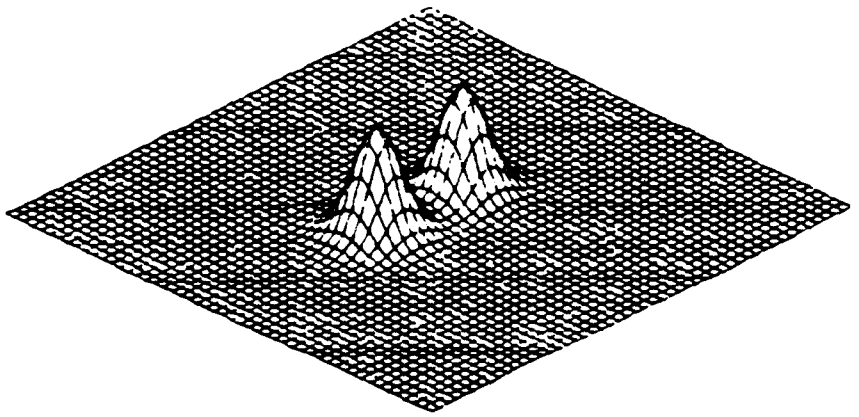
Actual intensity distribution



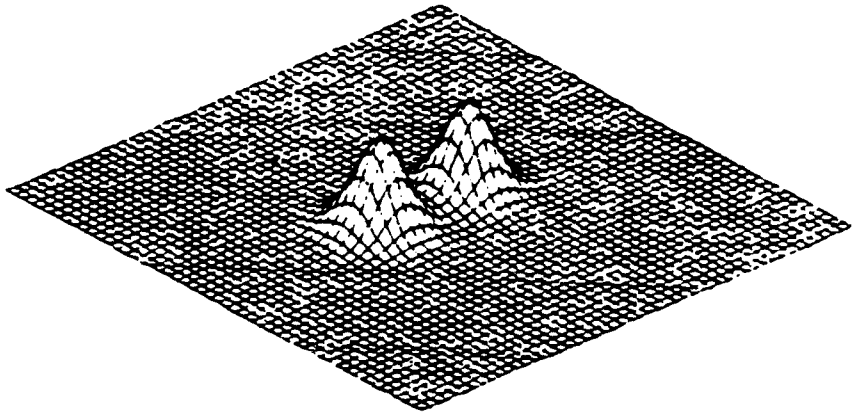
Predicted intensity distribution, $t^{*}=.108$

Figure 15

Predicted laser intensity distributions with spike widths of six plate thicknesses.



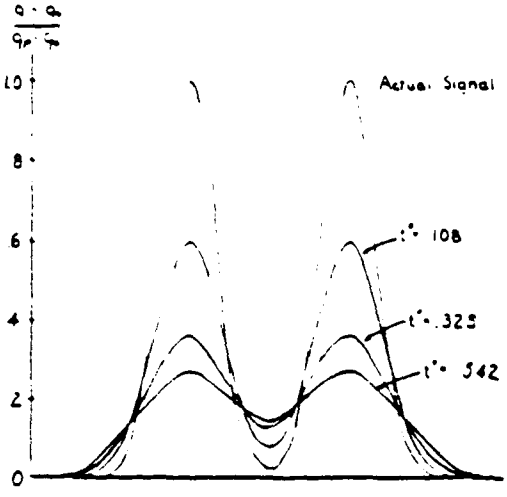
Predicted intensity distribution $t^* = .325$



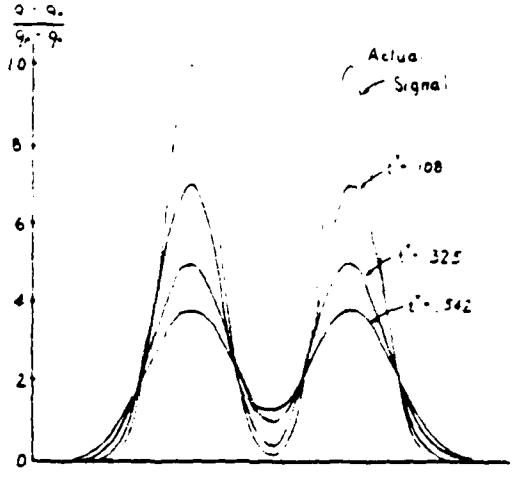
Predicted intensity distribution, $t^* = .542$

Figure 16

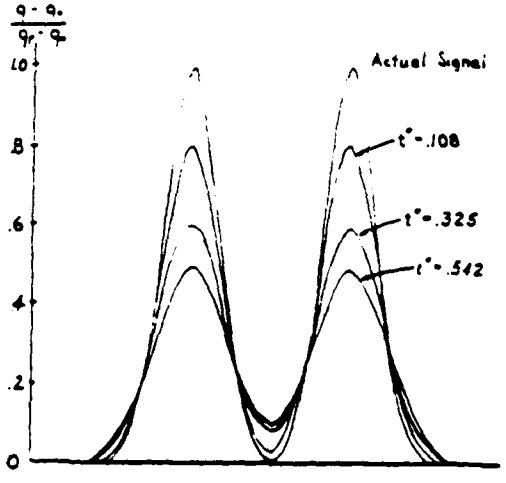
Predicted laser intensity distributions with spike widths of six plate thicknesses.



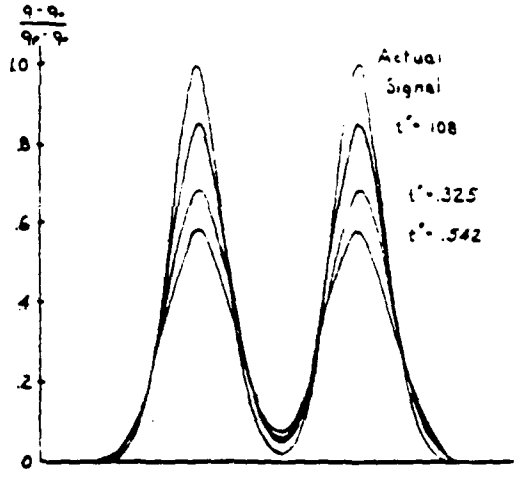
Spike width is three plate thicknesses



Spike width is four plate thicknesses



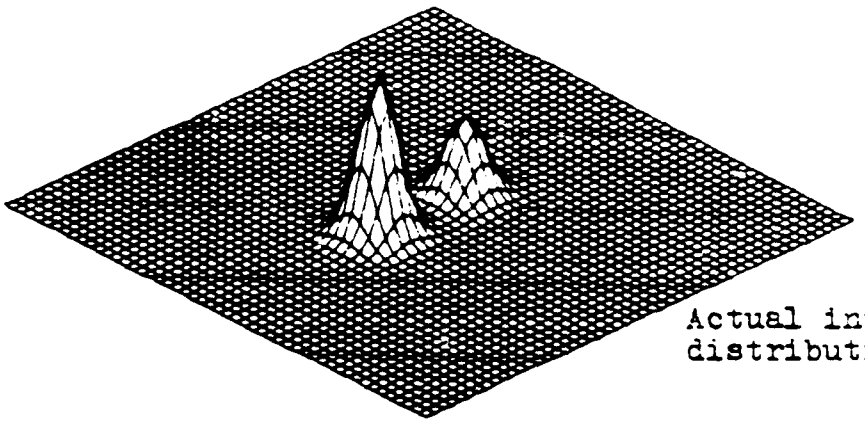
Spike width is five plate thicknesses



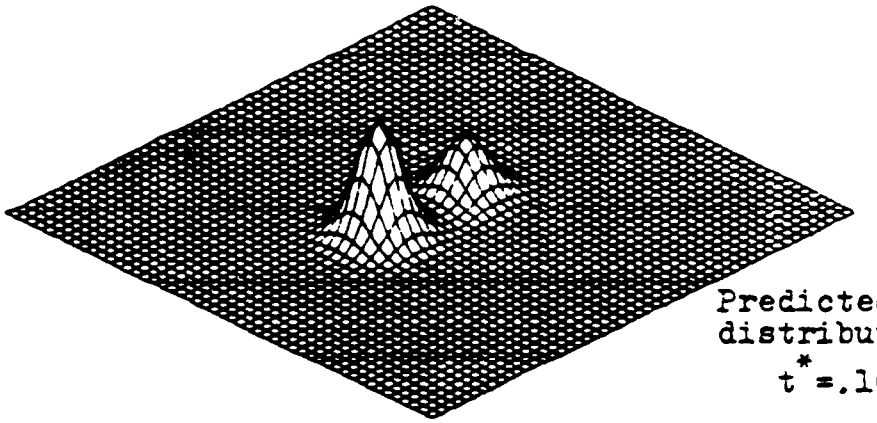
Spike width is six plate thicknesses

Figure 17

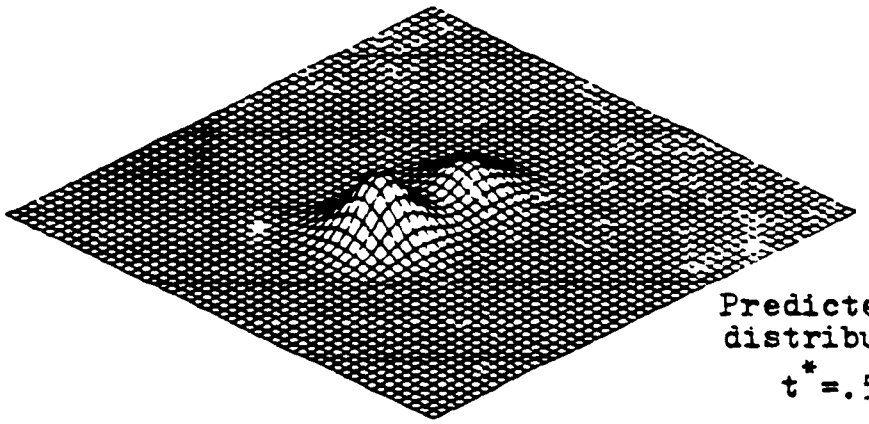
Predicted intensity profiles of distributions with two adjacent spikes.



Actual intensity distribution



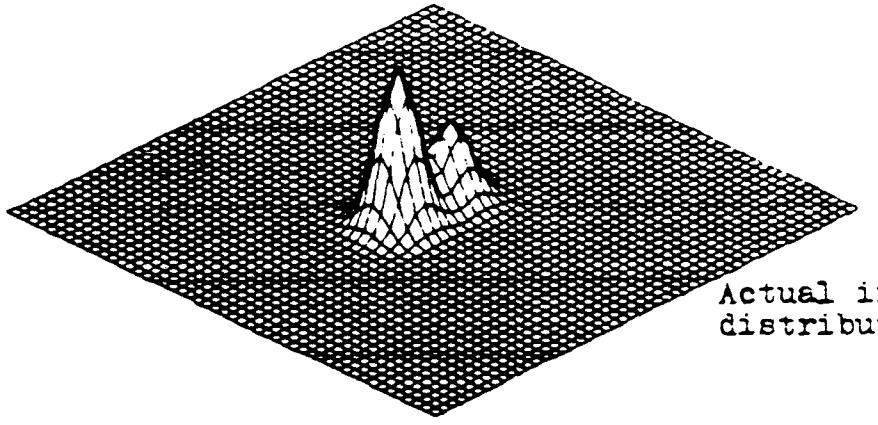
Predicted intensity distribution, $t^* = .108$



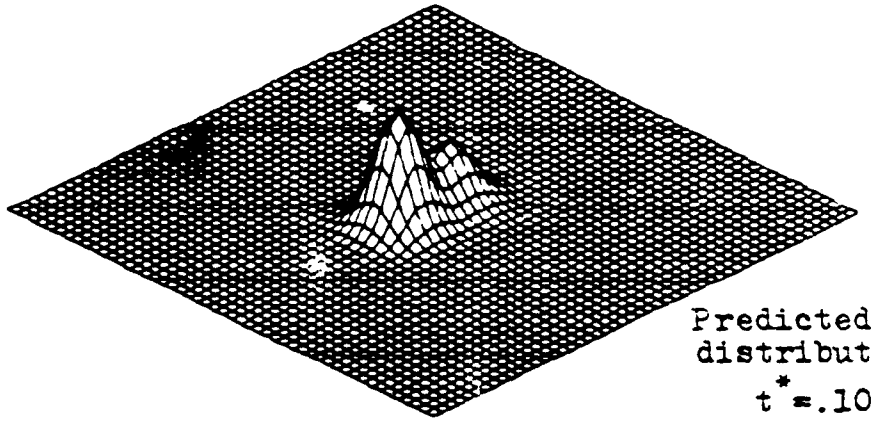
Predicted intensity distribution, $t^* = .542$

Figure 18

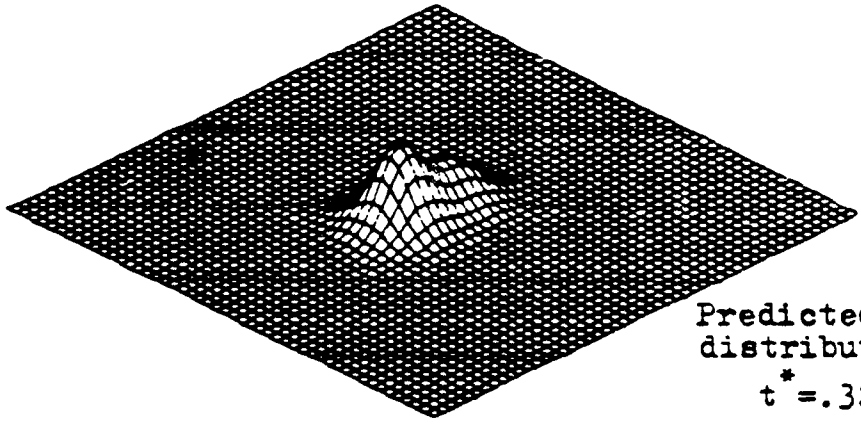
Predicted intensity distributions with two spikes with widths of four plate thicknesses.



Actual intensity distribution



Predicted intensity distribution, $t^* = .108$



Predicted intensity distribution, $t^* = .325$

Figure 19

Predicted intensity distributions with two spikes with widths of four plate thicknesses.

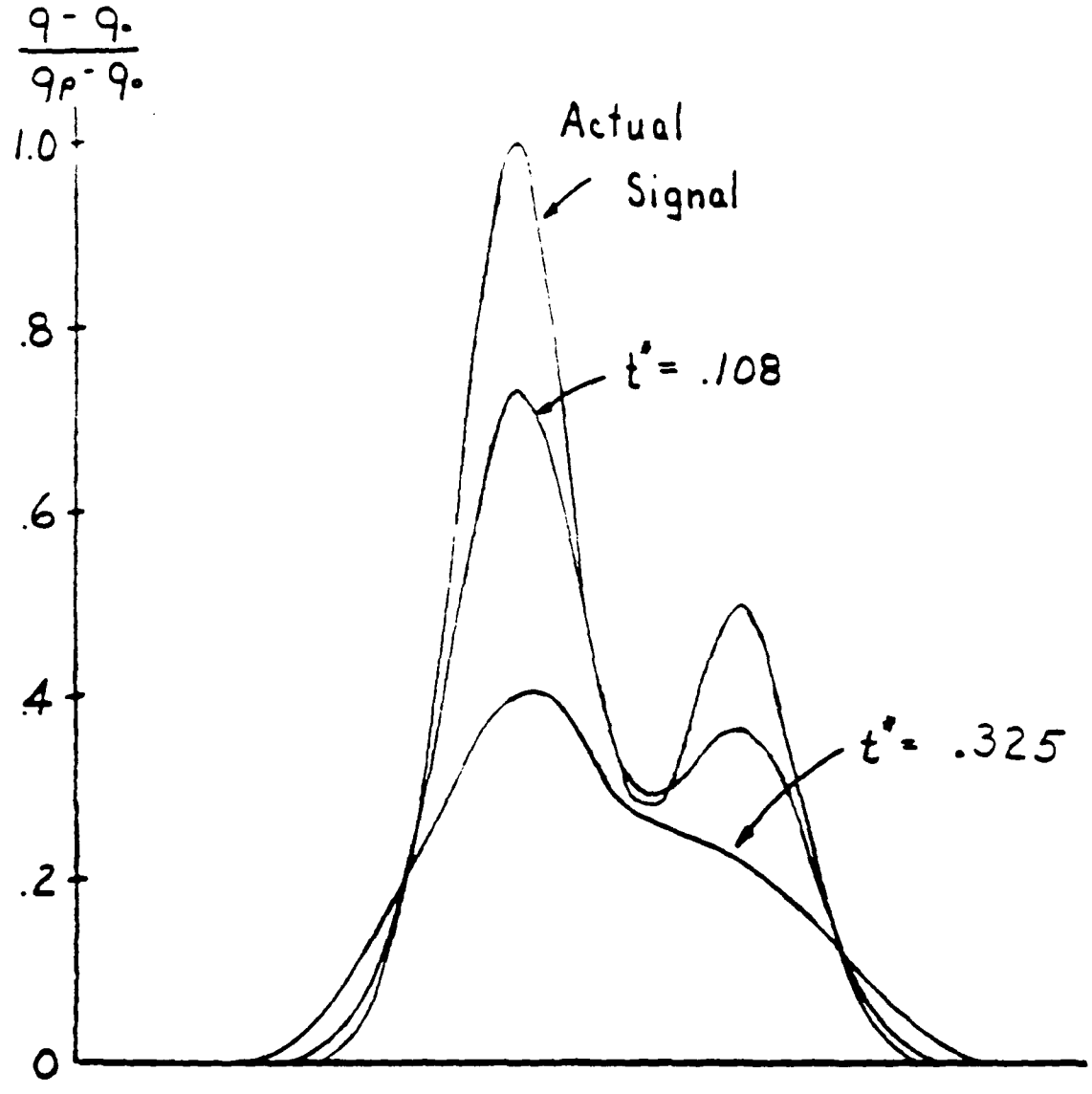
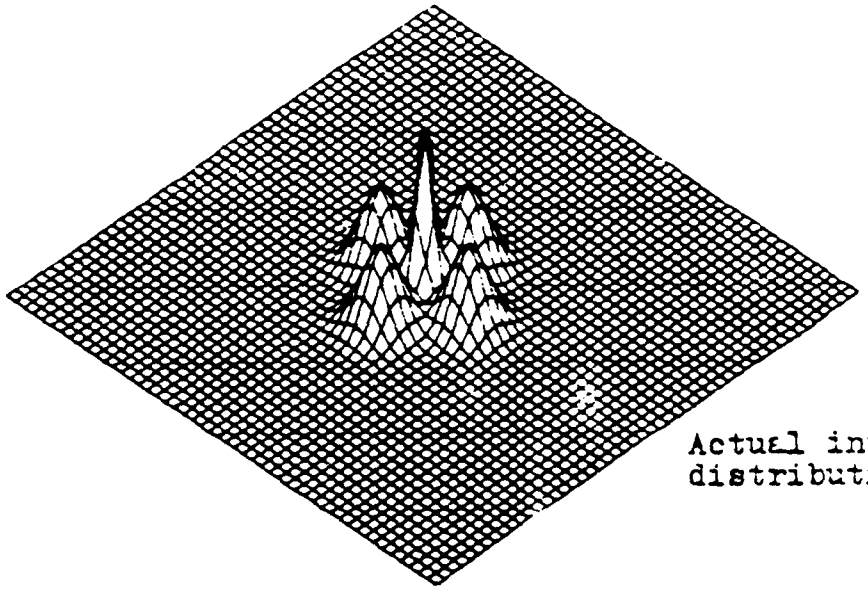
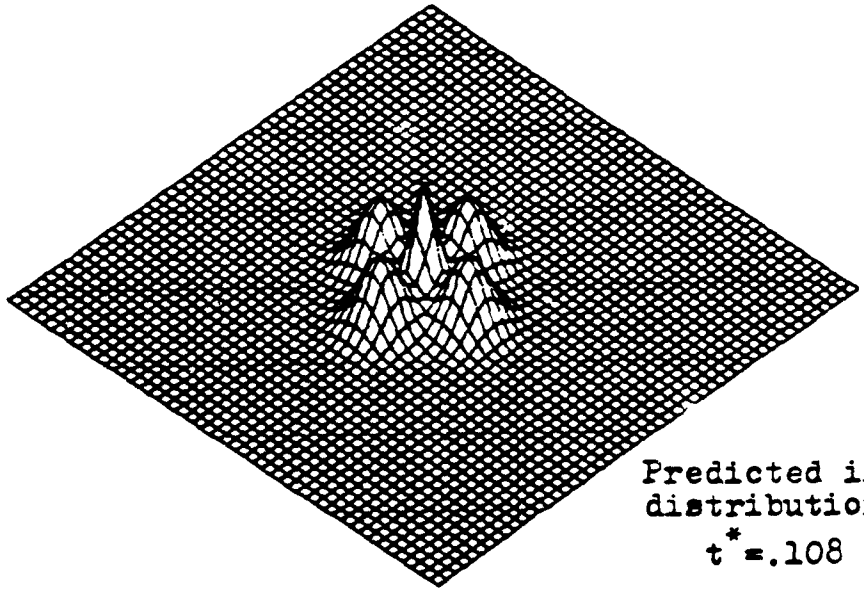


Figure 20

Predicted intensity profiles for adjacent spikes with widths of four plate thicknesses.



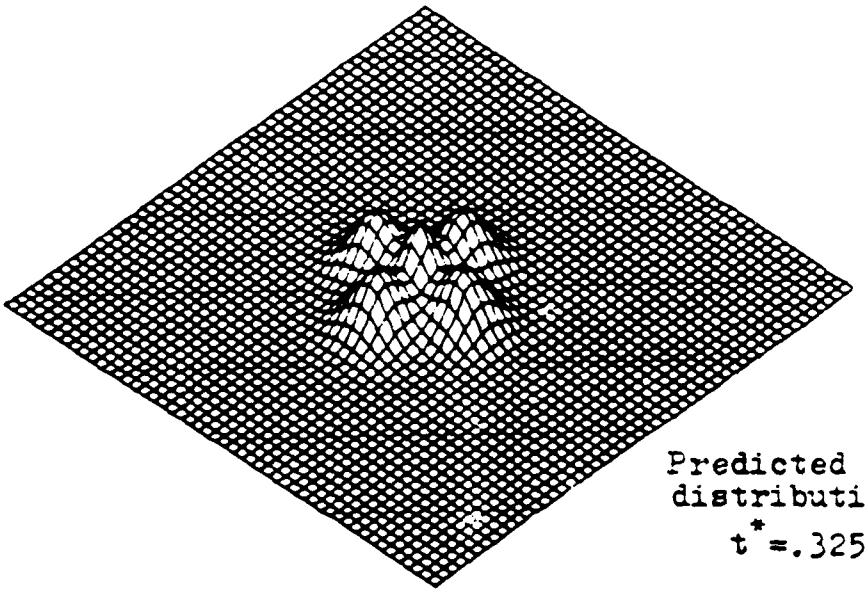
Actual intensity distribution



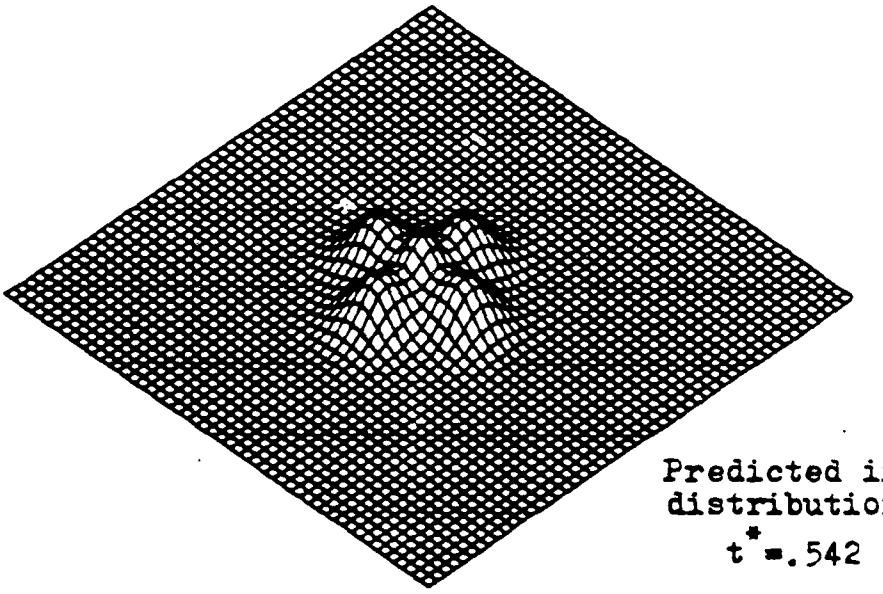
Predicted intensity distribution, $t^* = .108$

Figure 21

Predicted intensity distributions with five spikes. The central spike has a width of three plate thicknesses, and the other four have widths of six plate thicknesses.



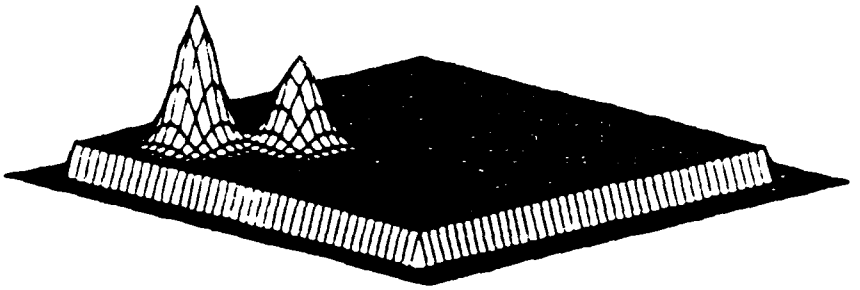
Predicted intensity distribution,
$t^* = .325$



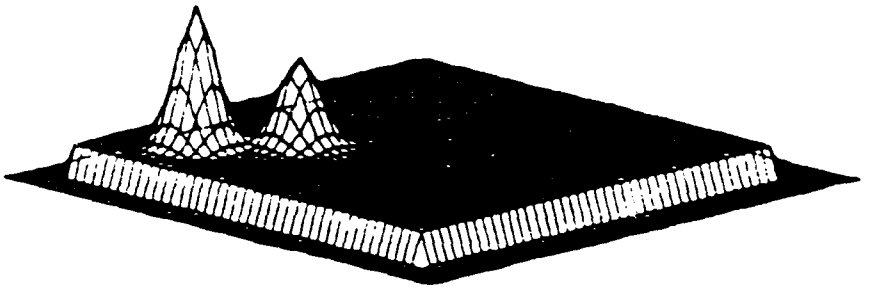
Predicted intensity distribution,
$t^* = .542$

Figure 22

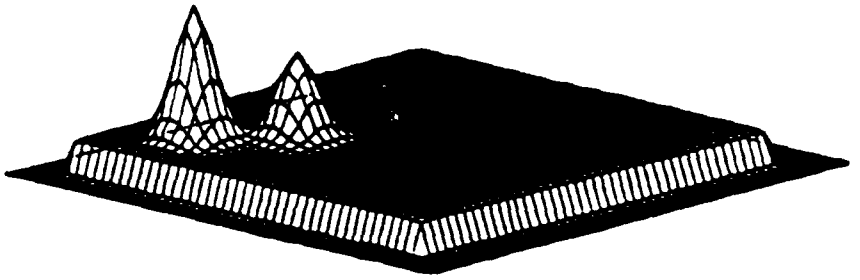
Predicted intensity distributions with five spikes. The central spike has a width of three plate thicknesses, and the other four have widths of six plate thicknesses.



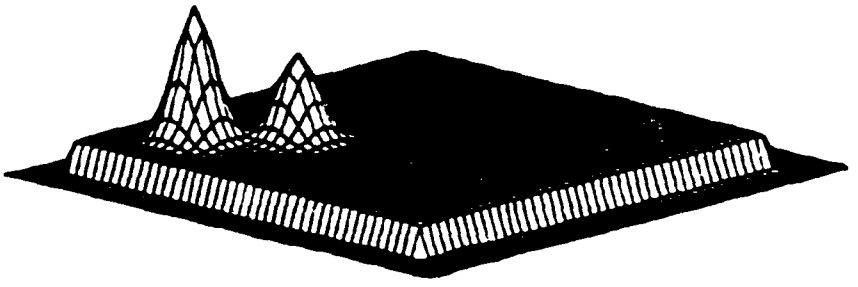
Actual intensity distribution



Predicted intensity distribution, $t^* = .108$



Predicted intensity distribution, $t^* = .325$



Predicted intensity distribution, $t^* = .542$

Figure 23

Predicted intensity distributions for a square laser beam with a width of 20 cm.

is the same as for the previous cases. This case was suggested by the Air Force Weapons Laboratory.

DISCUSSION AND CONCLUSIONS

The results for two adjacent intensity spikes with widths of two to six plate thicknesses are consistent with the results of Shadday (5) and (6), shown in figure (6). The narrowest spikes are most poorly resolved. For the case with spike widths of two plate thicknesses, there is considerable filling in of the valley separating the two spikes, and after an elapsed time of $t^* = .325$ (.3 seconds), the predicted intensity distribution is a single long hump, barely discernable as two distinct spikes.

The predicted intensity distributions for two adjacent spikes with some overlap, figures (19) and (20), show considerable filling in of the region between the two spikes. The two spikes have widths of four plate thicknesses, wide enough for them to be independently reasonably well resolved, but the less intense of the two spikes is not discernable as a separate spike after $t^* = .325$, (.3 seconds).

In figures (21) and (22), the predicted intensity distributions for a cluster of five spikes, the central one narrow and intense, are shown. As one might expect, the capability to resolve the central narrow spike drops off quickly, and after $t^* = .325$, the predicted intensity distribution is one with five spikes of essentially equal intensity.

The last case shown is for a laser beam with a square cross-section and two spikes in one corner of the beam. This distribution was suggested by the Air Force Weapons Laboratory, and the predicted intensity distributions faithfully reproduce the actual intensity distribution on the front surface of the target plate. The two intensity spikes have widths of 16.7 plate thicknesses, and the target plate laser intensity measurement technique can accurately measure laser beam intensity distributions with large scale features. Narrow intensity spikes are the features that are difficult to resolve.

The variety of laser beam intensity distributions that could be run is endless, but the results of the several cases that were run point out limitations of the target plate measurement technique. Intensity spikes narrower than four plate thicknesses are poorly resolved. Multi-dimensional conduction effects can mask a spike with a low peak intensity, if it is close to a spike with a high peak intensity. Large scale features in an intensity distribution are faithfully reproduced. The numerical results of this model vividly illustrate the benefit of collecting data from the rear surface of the target plate quickly. As the elapsed exposure time of the target plate to the laser beam increases, the importance of three-dimensional conduction effects increases.

The three-dimensional numerical model is a useful diagnostic tool for the target plate laser intensity measurement technique. The predicted laser beam intensity distributions are the theoretical best that can be measured at a specified time, since they are based on the complete spatial temperature distribution on the rear surface of the target plate at the specified time. This is more complete data than is collected from the response of the temperature sensitive paint.

ACKNOWLEDGMENT

This research was sponsored by the Air Force Office of Scientific Services, Bolling AFB, DC. Contract number F49620-85-C-0013/SB5851-0360. The author is grateful for the support.

REFERENCES

1. Lamar, C.R.,"Laser Diagnostics by Heat Conduction", Presented to the 1986 Joint AIAA/ASME Conference on Thermophysical Properties.

2. Ozisik, M.N., Heat Transfer A Basic Approach, McGraw-Hill Book Company, 1985.

3. Taylor, R.E., Larimore, J.,"Thermophysical Properties of Nickel and Stainless Steel 304", HTMIAC REPORT Air Force Weapons Laboratory, March 1986.

4. Roach, P.J., Computational Fluid Dynamics, Hermosa Publishers, Albuquerque, N.M., 1976.

5. Shadday, M.A.,"Two-Dimensional Conduction Effects in High Power CW Laser Target Plates", End of Summer Effort Report to the Air Force Weapons Laboratory, Kirtland AFB, N.M., August 1986.

6. Shadday, M.A., Couick, J.R.,"Two-Dimensional Thermal Conduction Effects in High Power CW Laser Target Plates", International Symposium on Thermal Problems in Space-Based Systems, HTD-Vol. 83, pp. 13-18, Presented at the 1987 ASME Winter Annual Meeting.

FINAL REPORT NUMBER 52
REPORT NOT AVAILABLE AT THIS TIME
Dr. William Wheless
760-7MG-068